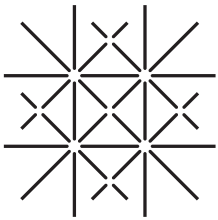
Verbesserung der Datenqualität von Online-Panels mittels Schätzung von Teilnahmewahrscheinlichkeiten

Dissertation zur Erlangung der Würde eines Doktors der Philosophie, vorgelegt der Philosophisch-Historischen Fakultät der Universität Basel

Gordon Wiegand

2011



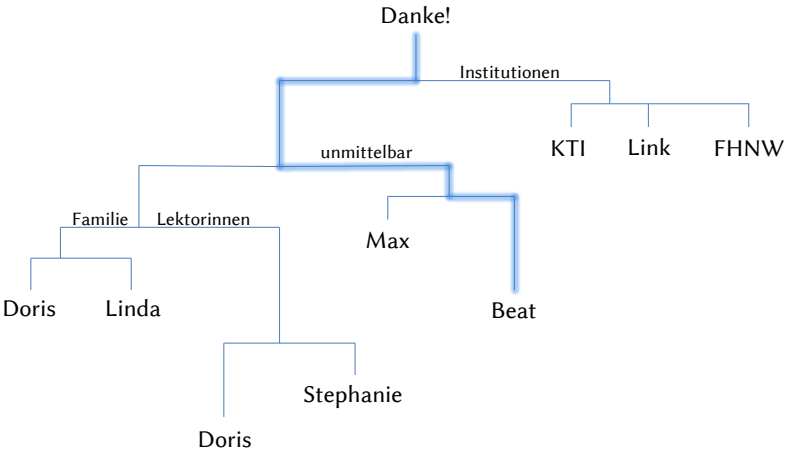
Genehmigt von der Philosophisch-Historischen Fakultät der Universität Basel auf Antrag von Prof. Dr. Max Bergman (Referent) und Prof. Dr. Beat Hulliger (Korreferent).

Basel, den 17. August 2011

Die Dekanin Prof. Dr. Claudia Opitz-Belakhal

Methodology, like sex, is better demonstrated than discussed,
though often better anticipated than experienced
Edward Leamer, 1983

Danksagung



Ich möchte mich herzlich bei allen bedanken, die an der Entstehung dieser Arbeit beteiligt gewesen sind. Zu allererst möchte ich **Beat Hulliger** für die grosse Hilfe und vielen Ideen danken! **Max Bergman** hat hilfreich verhindert, dass ich den soziologischen Kontext der Arbeit aus den Augen verliere und mich in unwichtigen Details verliere.

Die Arbeit ist im Rahmen des **KTI**-geförderten Projekts «Qualität von Online Panels» (Ref. Nr. 10092) als Kooperation der Fachhochschule Nordwestschweiz (Institute for Competitiveness and Communication, ICC) sowie dem **LINK**-Institut für Markt- und Sozialforschung entstanden. LINK ist besonders für die Realisierung der Befragung und die Adaption der Fragebögen zu danken. Daniela Lussmann danke ich für die Übersetzung des Fragebogens ins Französische.

Vielen Dank auch **Stephanie Greiwe** und **Doris Wiegand**, die so geduldig waren, die Arbeit von sehr vielen Rechtschreib- und Grammatikfehlerfehlern zu befreien.

Weiterhin danke ich allen bekannten und unbekannten Programmiererinnen und Programmierern der Open-Source-Community, die die Arbeit in Bezug auf die Programmierung der Auswertung und auch des Schriftsatzes erst ermöglicht haben. Ohne sie wäre die Arbeit in dieser Form nicht möglich gewesen.

Natürlich möchte ich mich auch bei meiner Frau Doris und meiner Tochter **Linda** für ihre Geduld und Unterstützung bedanken. Nicht zuletzt danke ich auch meinen Eltern **Werner** und **Carola Wiegand**.

Vorwort

Die Struktur des Dokumentes ist zwar im Inhaltsverzeichnis abgebildet, trotzdem mag es helfen, nochmals kurz die Struktur zu erläutern, die in Abbildung 0.1 illustriert ist. Eine inhaltliche Begündung für die Struktur wird noch in der Einleitung (Kap. 1) gegeben werden. Trotzdem soll die Abbildung nochmals der Orientierung dienen. Die fünf wichtigen Abschnitte der Arbeit sind in der Abbildung auf der linken Seite. Das Verfahren, mit dessen Hilfe eine Bias-Reduktion bei Web-Panel-Befragungen erreicht werden soll heisst PSA und wird in 3 eingeführt. Es schliesst sich die Beschreibung der experimentellen Befragung in Kap. 5 an, mit deren Hilfe die Methode spezifiziert und überprüft werden soll. Teil der Methode ist eine Modellierung, die in Kap. 7 beschrieben wird. Der eigentliche Test der Methode wird in Kap. 9 durchgeführt. Da sich zeigt, dass die Methode nicht zu dem gewünschten Erfolg führt, wird im abschliessenden Kap. 10 nochmals überprüft, ob die Methode selbst mangelhaft ist, oder ihre Spezifikation.

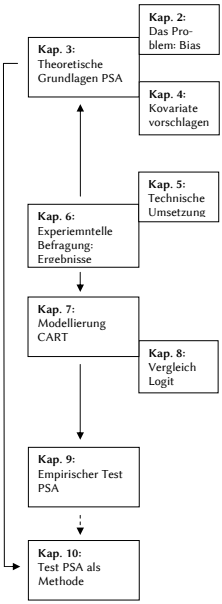


Abbildung 0.1: Struktur des Dokuments

In der elektronischen Version dieses Dokuments werden alle Elemente, die verlinkt sind, umrahmt. Dazu gehören URLs, Links zu anderen Kapiteln, Tabellen, Formeln sowie zu Literaturreferenzen in der Bibliografie. Den meisten Kapiteln ist eine kurze Einleitung vorangestellt, die etwas eingerückt wurde und mit ► . . .◄ umrahmt wird. Sie

soll jeweils helfen, den «Roten Faden» nicht zu verlieren und einen kurzen Überblick über das Kapitel zu geben. Ausserdem sind in insbesondere den Theorie-Kapiteln Randnotizen eingefügt worden, die die Orientierung im Text weiterhin erleichtern helfen sollen.

Inhaltsverzeichnis

1 Einführung

1.1 Die Bedeutung und Problematik von Web-Panel-Befragungen

Fehler bei Befragungen

Die Qualität von Befragungen hängt von vielen Faktoren ab. Groves et al. (2004) systematisieren Fehler, die zu Verzerrungen bei Schätzungen führen können, unter der Überschrift *Total Survey Error*[1]. Vereinfacht gesagt, gibt es zwei Quellen von Fehlern bei Befragungen. Eine Gruppe von Fehlern betrifft die Durchführung der Befragung selbst und umfasst neben vielen anderen Dingen Aspekte wie die korrekte Frageformulierung, die Qualität des Fragebogens insgesamt und die Schulung der Interviewer. Dieses Themenfeld soll nicht Thema dieser Arbeit sein, obgleich es in seiner Bedeutung keinesfalls unterschätzt werden darf.

Ein zweiter Strang weist auf mögliche Fehlerquellen bezüglich der Repräsentativität von Befragungen hin. In diesem Strang werden all diejenigen Faktoren aufgelistet, die zu einem *Bias* führen können, dessen Quelle z. B. der systematische Ausfall von Befragten sein kann. Zu diesem Strang gehört auch die Datenaufbereitung und Analyse.

Um einen Bias zu reduzieren, sind prinzipiell zwei verschiedene Strategien denkbar (Stoop, 2005). Zunächst kann man ganz praktisch orientiert versuchen, so viele Angefragte wie möglich zu motivieren, tatsächlich an der Befragung teilzunehmen. Es gibt eine ganze Reihe von Möglichkeiten, dies zu erreichen[2]. Weiterhin bzw. ergänzend ist es möglich, einen entstandenen Bias nachträglich zu korrigieren. Die vorliegende Arbeit wird sich auf diesen Aspekt der Qualitätsicherung konzentrieren.

Ziel der Arbeit

Die Arbeit möchte exemplarisch überprüfen, ob die Qualität der Daten aus Online-Panel Befragungen mittels *propensity score adjustment* (PSA)

1 Ähnliche Systematiken schlagen auch andere vor, z. B. Bethlehem (2010).

2 Für eine Übersicht siehe neben Stoop (2005) (sehr gut), Bethlehem (2002) (auch sehr gut, technischer), de Vaus (2002) (vier Bände) insbesondere auch Groves et al. (2002) (Standardwerk) und Biemer und Lyberg (2003) (sehr informativ) und Groves und Couper (1998) (meistzitiert).

verbessert werden kann. Die mögliche Verbesserung soll anhand des Online-Panels des LINK Institut für Markt- und Sozialforschung gezeigt werden.

> **Bemerkung:** Das Panel von LINK ist besonders geeignet, da LINK laufend neue Panelisten[3] aktiv rekrutiert. Dies ermöglicht insbesondere die experimentelle Modifikation des Rekrutierungsprozesses. Der Anspruch von LINK, eine möglichst hohe Datenqualität zu erreichen, war grundlegend für die Etablierung der Zusammenarbeit. Das Panel wird nicht nur bei typischen Marktforschungsstudien eingesetzt, sondern auch bei vielen wissenschaftlichen Studien im Auftrag von Universitäten, Behörden und anderen Forschungsträgern.

Bedeutung von Web-Befragungen

Befragungen sind nach wie vor die wichtigste Methode der Datensammlung in den Sozialwissenschaften. Die Befragungslandschaft hat sich allerdings in den letzten Jahren stetig verändert. Eine aktuelle Übersicht findet sich z. B. bei Tortora (2009) oder bei Bethlehem (2010). Es lässt sich zusammenfassen, dass Befragungen über das Internet zunehmend andere Befragungsmodi wie persönliche Interviews (z. B. CAPI), Telefoninterviews (CATI) und schriftliche Befragungen (CASI, CASAQ) ersetzen[4]. Ausführlich zeigen dies für den Bereich der Marktforschung Postoaca (2006), schon früh für die Sozialforschung im Allgemeinen Couper et al. (1998), für die Psychologie Göritz (2007) und für die Sozialmedizin Couper (2007).

Der Bereich, in dem Befragungen mittels Web-Panel am häufigsten durchgeführt werden, ist sicherlich die Marktforschung (Callegaro und Disogra, 2008). Gemäss
Comley (2007) wurden in den USA und in Europa rund ein Drittel aller Befragungen in der Marktforschung mittels Web-Panel durchgeführt. In der Schweiz beträgt der Anteil der Web-Befragungen, die durch durch den Verband Schweizer Markt- und Sozialforscher, VSMS, durchgeführt wurden,

3 Alle Bezeichnungen von Personen in dieser Arbeit sind geschlechtsneutral gemeint. Aufgrund der besseren Lesbarkeit wurde auf die Aufführung beider grammatikalischen Geschlechtsbezeichnungen verzichtet. Die grammatikalisch männliche Form wird verwendet, da sie meistens kürzer und gebräuchlicher ist und selbstverständlich nicht, um ein Geschlecht im nichtgrammatikalischen Sinne zu bevorzugen bzw. zu benachteiligen!

4 *CAPI* Computer-assisted personal interviewing
CATI Computer-assisted telephone interviewing
CASI Computer-assisted self interviewing
CASAQ Computer-administered self-administered questionnaires

im Jahr 2007 18 % (VSMS, 2009). Die Tendenz ist stark steigend. 2006 hat der Anteil noch 12 % betragen, 2005 8 %, 2004 4 % und im Jahr 2000 1 %. Davor wurden keine Befragungen via Internet in der Marktforschung durchgeführt[5]. Seit 2008 wird die Bedeutung der Befragungsmethoden nur noch bezüglich des Beitrags zum Gesamtumsatz verglichen und nicht mehr als Anteil an der Menge der Interviews. Hat der Anteil von Web-Befragungen am Umsatz der Branche 2008 noch 12 % betragen, ist er 2009 auf 16 % gestiegen (VSMS, 2010).

Der wachsende Anteil und damit die zunehmende Bedeutung von Web-Befragungen an allen Befragungen im Bereich der Markt- und Sozialforschung kann auf zwei wichtige Vorteile zurückgeführt werden. Der Vorteil für Auftraggeber solcher Studien liegt in der hohen Geschwindigkeit. Sowohl die Erstellung, der Rücklauf als auch die ersten Auswertungen benötigen wenig Zeit. Schon innerhalb weniger Tage nach Auftragsvergabe können erste Ergebnisse zur Verfügung stehen. Bei manchen Fragestellungen, wie z. B. bei Wirkungsstudien von Werbung, ist die hohe Rücklaufgeschwindigkeit sogar eine essentielle Voraussetzung.

Auf der Seite der Anbieter liegt der Vorteil in den relativ geringen (variablen) Kosten. Eine Schulung von Interviewerpersonal wie bei telefonischen und persönlichen Interviews (CATI bzw. CAI) ist nicht notwendig und auch sonst ist die Administration der Befragung vergleichsweise unkompliziert. Der postalische Versand von Fragebögen sowie die anschliessende aufwändige elektronische Erfassung des Rücklaufs entfallen. Moderne Befragungstools machen die Erstellung und technische Abwicklung der Onlinebefragungen unkompliziert und ermöglichen die leichte Einbindung von Multimediaelementen wie Tonsequenzen, Bildern und Videos. Web-Befragungen sind also vergleichsweise günstig.

Probleme von Web-Befragungen

Aber auch zwanzig Jahre nach ihrer Einführung sind Befragungen über das Internet trotz dieser Vorteile noch immer nicht unumstritten (Reips, 2007; Batinic, 2002; Fricker und Schonlau, 2002, , und viele andere). Zu Recht wird häufig die mangelhafte Qualität der resultierenden Daten auf-

5 Alle Anteilszahlen sollen nur die Tendenz verdeutlichen. Es ist durch den VSMS nicht klar dokumentiert, wie die Anteile berechnet wurden. Die Angaben beziehen sich nur auf Institute, die dem VSMS angeschlossen sind. Diese Institute führen eine Vielzahl von Befragungen durch, insbesondere auch für die Forschung, die dadurch vermutlich auch Teil der Berechnung sind. International vergleichbare Zahlen zu den Anteilen an Web-Befragungen existieren nicht.

grund methodischer Mängel kritisiert, wie z. B. zuletzt durch Rhall und Fine (2008) und Scherpenzeel (2008). Beispielsweise ist es nicht einfach, alle Teile der Bevölkerung mit Hilfe des Internets mit vergleichbarer Wahrscheinlichkeit zu erreichen.

Sicherlich ist die Qualität von Daten aus Befragungen allgemein ein Problem in der Sozialforschung. Es gibt sehr viele Quellen von Fehlern, die häufig nicht kontrolliert werden können (Groves, 1989). Nicht zuletzt wegen der Schwierigkeit, valide Daten zu bekommen, zählen die Sozialwissenschaften zu den «weichen» Wissenschaften in Abgrenzung zu den «harten» (Popper, 1935; Lakatos, 1995). Auch deshalb zählt es zu den dringlichen Aufgaben der Sozialwissenschaften und der Soziologie im Speziellen, ihren Analysen Daten möglichst hoher Qualität zu Grunde zu legen.

Auch wegen der steigenden Bedeutung von Web-Befragungen liegt der Fokus dieser Arbeit bei dieser Befragungsmethode, speziell bei Befragungen mittels Panels. Die vorgestellte Methode der Datenverbesserung, das *propensity score adjustment*, lässt sich auch auf andere Befragungsmodi anwenden. Einen Einblick hierzu gibt Cobben (2009).

Zunächst scheinen Befragungen über das Internet mit den anderen Befragungsformen viel gemein zu haben; nur die technische Umsetzung des Fragebogens ist verschieden. Es gibt aber (mindestens) zwei Probleme, die sich bei Web-Befragungen noch stärker als bei den anderen Modi manifestieren: *under-coverage* und *self-selection*[6]. *Under-coverage* bedeutet, dass es mit dem gewählten Auswahlmechanismus einer Stichprobe nicht möglich ist, alle Elemente einer Population zu erreichen. Insbesondere ist es häufig so, dass die angestrebte Grundgesamtheit grösser ist als die Menge der regelmässigen Internetnutzer. *Under-coverage* heisst also, dass es eine Gruppe von Elementen der Grundgesamtheit gibt, über die mittels Stichprobe zwar eine Aussage getroffen werden soll, diese aber prinzipiell aus der Stichprobe ausgeschlossen ist[7]

Under-coverage I: Internet-Penetration

Das Problem des *under-coverage* bei Web-Befragungen wird zwar immer kleiner, bleibt aber weiterhin bestehen. Abbildung 1.1 zeigt die Entwicklung

6 Vgl. neben den meisten bereits zitierten Autoren exemplarisch Bethlehem (2010)

7 Selbstverständlich gibt es dieses Phänomen auch bei anderen Befragungsmodi. Z. B. werden häufig bei CATI-Befragungen nur Festnetzanschlüsse angerufen. Insebesondere bei Jugendlichen ist es allerdings zunehmend verbreitet, nur ein Mobiltelefon zu besitzen, siehe dazu weiter unten.

der Internet-Penetration in der Schweiz seit 1998[8]. Mit *ENK* ist der engere Nutzerkreis gemeint. Gemäss Definition des BFS sind dies alle Personen, die das Internet mehrmals pro Woche verwenden. Mit *WNK*, dem weiteren Nutzerkreis, sind alle Schweizer gemeint, die das Internet mindestens einmal in den vergangenen sechs Monaten verwendet haben. Aktuell (Stand 9. März 2010) gehören zum ENK 74 % der Schweizer Bevölkerung, zum WNK 82.1 %.

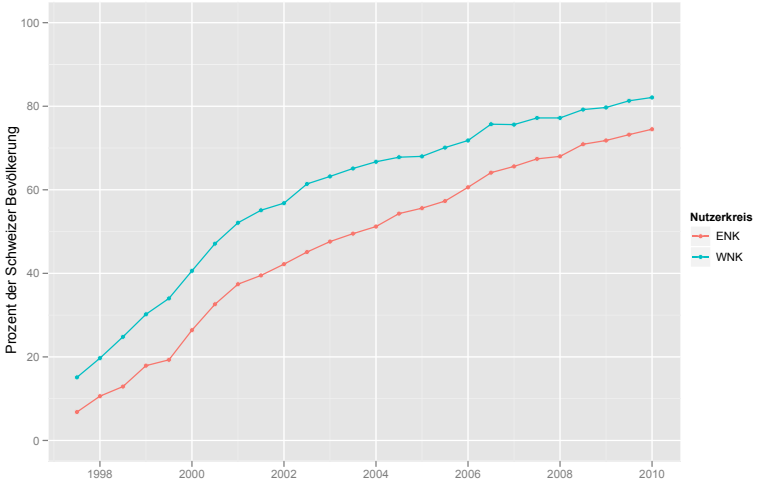


Abbildung 1.1: Internet-Penetration in der Schweiz

Der Trend zu einer immer höheren Internetpenetration ist nebenbei bemerkt gegenläufig zum Trend bei der Erreichbarkeit von Personen und Haushalten via Festnetz, dem zur Zeit wichtigsten Befragungsgerät. Das Maximum wurde 1995 mit 62.1 Zugangsleitungen pro 100 Personen erreicht.

8 Quelle der Daten: BFS, NET-Metrix-Base, `www.bfs.admin.ch/bfs/portal/de/index/themen/16/03/key/ind16.Document.25576.xls`, Stand 01.10.2010

Mittlerweile sind es nur noch 47.3 (Stand 2008)[9]. Dies liegt insbesondere an der stetig wachsenden Substitution von Festnetzanschlüssen durch Mobilfunktelefone. Mittlerweile kommen auf 100 Schweizer 115.5 Mobiltelefone. Befragungen mittels Mobiltelefonen haben sich aber bisher nicht etablieren können, auch da es sehr aufwändig ist, aus allen Mobiltelefonbesitzern eine Zufallsstichprobe zu ziehen. Ein Pendant zum Random-Digit-Dialing (RDD) Verfahren existiert zwar[10], wird aber nur wenig genutzt. Auch dieser Trend spricht dafür, dass die Bedeutung und Akzeptanz von Web-Befragungen zukünftig steigen wird.

> **Exkurs Internetpenetration bei Schweizer Haushalten** 77 % der Haushalte verfügen aktuell über einen eigenen Internetzugang (Bundesamt für Statistik, 2011). Das BFS glaubt, dass das Potential der Erschliessung damit ausgeschöpft ist, da 20 % der Haushalte explizit zu Hause keinen Internetanschluss möchten, da sie keinen Bedarf haben, sich selbst einen Mangel an Kompetenzen attestieren oder körperliche Schwierigkeiten wie z. B. eine Behinderung haben. Das Vorhandensein eines Internetanschlusses variiert stark mit der Zusammensetzung des Haushaltes. Solche, in denen die älteste Person unter 50 Jahren alt ist, haben zu 95 % einen Anschluss; solche, deren ältestes Mitglied 70 Jahre oder älter ist, lediglich zu 33 %. Kinder spielen eine wichtige Rolle bei der Übernahme dieser (und auch anderer) Technologie. Nach Regionen betrachtet hat das Tessin mit 64 % angeschlossener Haushalte den geringsten Anteil angeschlossener Haushalte.

Under-coverage II: fehlende Antwortbereitschaft

Auch wenn die Internet-Penetration vollständig wäre, es also kein *undercoverage* gäbe, bleibt das Problem fehlender Antwortbereitschaft, welches auch zu einem Bias führen kann. Obwohl die Bereitschaft, an Web-Befragungen teilzunehmen, oft relativ hoch ist (Batinic (2002) und LINK berichten überein-

9 Quelle BFS und BAKOM, `www.bfs.admin.ch/bfs/portal/de/index/themen/16/04/key/approche_globale.Document.25540.xls`, letzter Zugriff 1. Nov 2010. Mit Festnetzanschlüssen sind sowohl PSTN und ISDN Anschlüsse gemeint.

10 Das Verfahren wurde Gabler und Häder (2007) für das BFS entwickelt.

stimmend von rund 60 % Respondenten[11]) bedeutet dies noch nicht, dass die Bereitschaft auch zukünftig so hoch bestehen bleibt. Man kann vermuten, dass sie wegen zunehmender Anzahl von Befragungen und zunehmendem SPAM zukünftig sinken werden. Ein nicht-zufälliger Teil der Bevölkerung wird also möglicherweise eine sinkende Teilnahmebereitschaft haben.

Web-Panels

Under-coverage und Nonresponse führen (meistens) zu einem systematischen Ausfall von Befragten und damit zu einem Bias, also häufig zu gravierenden Fehlern. Ein erstes Mittel, das Problem zumindest abzumildern, ist es, Befragungen via Internet nicht unmittelbar durchzuführen, sondern via Panels zu organisieren. Bei Panels werden dieselben Befragten wiederholt zu Befragungen eingeladen. Es lohnt sich daher eher, für die Rekrutierung der Befragten einen hohen Aufwand zu betreiben. Ausserdem ist es mit zunehmender Grösse des Panels[12] einfacher, gewisse Variablen zu kontrollieren und so z. B. geschichtete Stichproben für die einzelnen Befragungen zu ziehen oder Quotenverfahren anzuwenden. Ausserdem kann versucht werden, schon bei der Rekrutierung einigen Ausfallmechanismen entgegenzusteuern. Es ist beispielsweise möglich, «einfache» Ausfallmechanismen wie der gehäufte Ausfall einzelner Regionen oder Altergruppesgruppen auszugleichen[13]. Das vorrangige Ziel dieser Arbeit ist es, zu untersuchen, ob das *porpensity score adjustment (PSA)* im Kontext von Web-Panel Befragungen geeignet ist, den zu erwartenden Bias zu reduzieren. Dabei soll der Bias, der als Differenz zwischen Panelisten und der schweizer Wohnbevölkerung entsteht behandelt werden. Marktforschungsinstitute beziehen sich bei ihren Befragungen mit Web-Panels allerdings typischerweise auf die sogenannte Internetpopulation als Grundgesamtheit und nicht auf die Wohnbevölkerung. Über die Internetpopulation sind allerdings keine allgemeinen und zuverlässigen Aussagen bekannt; dies scheitert schon daran, dass sie schlecht zu definieren ist.

11 Die Höhe der Teilnahmerate ist selbstverständlich ein stark schwankendes Phänomen. Die Höhe ist abhängig von sehr vielen Dingen, die weiter unten besprochen werden (Kapitel 4, ab S. 55). Ein wichtiges Kriterium bei Panelbefragungen ist zum Beispiel die Panelpflege. Werden Befragte, die in Folge auf Einladungen nicht reagiert haben aus dem Panel eliminiert, steigen die berichteten Teilnahmeraten. Zu den Willkürlichkeiten bei der Berechnung von Teilnahmeraten siehe auch Abschnitt 3.5.1 ab S. 44. So überrascht es nicht, dass in der Literatur teilweise sehr unterschiedliche Teilnahmequoten berichtet werden (Scherpenzeel und Bethlehem, 2010). Viele Autoren berichten auch sehr viel niedrigere Teilnahmequoten.

12 Das Panel von Link hat zum Zeitpunkt dieser Arbeit eine Grösse von rund 110'000 Panelisten

13 Welchen Einfluss ein solches Vorgehen auf einen Bias der intererssierenden Variablen hat, is aber zunächst unbekannt.

Die Ausführungen beziehen sich nur auf Panels ohne *self-selection*. *Self-selection* bedeutet, dass die Auswahl der Befragten wenig bis gar nicht kontrolliert wird. Häufig werden Internet-Panels so organisiert, dass sich die Partizipienten selbst zur Teilnahme «einladen» können. Praktisch kann also jeder mitmachen, der ein Teilnahmeformular ausfüllt. Teilnehmende an Befragungen mit solchen Panels sind also de facto nur Personen, die einen Zugang zum Internet haben, auf die Seite mit dem Zugangsformular gestossen sind und bereit sind, an Befragungen teilzunehmen. Wichtigster Anreiz für eine solche Teilnahme sind wohl typischerweise die materiellen Anreize.

Vermutlich ist es (zumindest zur Zeit) ausgeschlossen, Daten aus Befragungen mit *self-selection* Zugang zu verbessern. Das Vorgehen ist zu weit von einer Zufallsauswahl entfernt, als dass sich noch etwas korrigieren liesse[14]. Die Arbeit fokussiert vielmehr auf Probleme, die aus *under-coverage* sowie durch modus-unabhängige aber biasbehaftete Ausfallmechanismen resultieren.

1.2 Forschungsdesign und Gliederung

Zusammenfassung

Das **propensity score adjustments** wird in Kap. 3 noch ausführlich eingeführt. Um das Verständnis zu verbessern, soll aber bereits hier ein kurzer Abriss der Methode gegeben werden.

PSA basiert auf der grundlegenden Annahme, dass die Teilnahmeneigung geschätzt werden kann. Aus dieser geschätzten Teilnahmeneigung können dann beispielsweise Gewichte abgeleitet werden, die eine Reduktion von Bias ermöglichen sollen. Die Gewichte werden so bestimmt, dass Personen mit einer geringen Teilnahmewahrscheinlichkeit ein relativ hohes Gewicht erhalten, um ihr «Fehlen» in der Stichprobe auszugleichen. Um die Teilnahmewahrscheinlichkeit schätzen zu können, sind so genannte Kovariate notwendig. Dabei handelt es sich um Variablen, die auf Individualebene sowohl für die Befragten, wie auch die Grundgesamtheit vorliegen müssen. Die Teilnahmeneigung kann dann bezüglich dieser Kovariate bestimmt werden. Wie gut die Biasreduktion ist, hängt insbesondere auch davon ab,

14 Dabei handelt es sich lediglich um eine sehr persönliche Einschätzung, die keinesfalls gesichert ist!

wie gut die ausgewählten Kovariate das tatsächliche Teilnahmeverhalten erklären können. Es ist ex ante nicht klar, welche Variablen dies möglichst gut zu leisten im Stande sind: dies muss vielmehr experimentell bestimmt werden. Kern dieser Arbeit ist der Versuch, solche Kovariaten mittels einer experimentellen Befragung zu identifizieren und aufbauend mittels eines geeigneten Modells die Teilnahmewahrscheinlichkeit zu schätzen.

Experimentelle Befragung

Um zu überprüfen, ob das PSA zu einer Verbesserung der Daten führen kann und um geeignete Kovariaten zu finden, wird eine experiemntelle Befragung durchgeführt. Diese besteht aus einer CATI-Befragung, die eine Reihe von Variablen abfragt, die potentiell die Teilnahmebereitschaft erklären können. In der CATI-Befragung ist die Frage nach der Bereitschaft an einem Web-Panel zu partizipieren eingeschlossen. Alle, die dem zugestimmt haben, haben später eine Einladung zu einer Befragung via Internet erhalten. Haben sie reagiert und den Fragebogen (der in dem Kontex dieser Arbeit irrelevant ist) beantwortet, wurden sie als Panelisten gezählt. Es ist so möglich, Panelisten mit Panelverweigerern bezüglich einiger Kontrollvariablen zu vergleichen. Abgeleitet kann die Wahrscheinlichkeit geschätzt werden, dass eine Person an einer Web-Befragung partizipiert. Einschränkend ist zu sagen, dass es sich nur um die bedingte Wahrscheinlichkeit der Teilnahme an einem Web-Panel handelt, nämlich unter der Bedingung, bereits an einer CATI-Befragung teilgenommen zu haben. Die echte Wahrscheinlichkeit der Teilnahme an einem Web-Panel ist mit dem hier gewählten Design nicht schätzbar. Die Umsetzung der Befragung selbst wird in Kapitel 5 nochmals detaillierter vorgestellt.

Kovariate

Um mit Hilfe der Teilnahmewahrscheinlichkeit geeignete Gewichte zur Reduktion von Bias ableiten zu können, ist es zunächst wichtig, das Teilnahmeverhalten in einem (statistischen) Modell beschreiben zu können. Grundlage für dieses Modell sind geeignete Fragen oder technisch ausgedrückt, geeignete Kovariaten zur Verfügung zu haben. Die empirischen Erfahrungen mit PSA bei Web-Panels ist bisher beschränkt. Es müssen daher Kandidaten für solche Kovariaten vorgeschlagen werden. Dazu werden in Kapitel 4 Theorien gesucht und besprochen, die das Potential haben, das Teilnahmeverhalten zu prognostizieren und die erlauben, empirisch prüfbare Indikatoren abzuleiten. Aufbauend auf diese Überlegungen, wird in diesem Kapitel auch der der experimentellen Befragung zu Grunde liegende Fragebogen abgeleitet.

Beschreibung des Teilnahmeverhaltens

Das eigentliche Ziel dieser Arbeit ist es zu überprüfen, ob eine Biasreduktion mit Hilfe der Schätzung von Teilnahmewahrscheinlichkeiten möglich ist. Ein eher implizites aber inhaltlich notwendiges Ziel ist es, das Teilnahmeverhalten zunächst beschreiben zu können. Es ist sicher interessant und hilfreich für das Verständnis des Teilnahmeverhaltens, wenn die Panelisten mit den Verweigerern bezüglich der erhobenen Kovariate verglichen werden können. Dies wird in Kapitel 6 strukturiert vorgenommen.

Modellierung

Neben der Erhebung geeigneter Kovariate ist es elementar, das Teilnahmeverhalten möglichst gut zu modellieren. In der Statistik sind verschiedene Methoden bekannt, die dies prinzipiell zu leisten vermögen. In den Kapiteln 7 und 8 werden zwei verschiedene Verfahren vorgeschlagen und ausprobiert. In Kapitel 7 wird das Teilnahmeverhalten mit Hilfe eines *regression tree*[15] (Baum-Modell) modelliert. Typischerweise verwendet man allerdings ein Lineares Modell, was im anschliessenden Kapitel 8 nachgeholt wird. Die Modellierung mittels Baum-Modell war notwendig, da es zu einer grossen Zahl fehlender Werte in der Befragung gekommen ist, deren Anzahl selbst wieder ein guter Indikator für die Partizipation am Panel ist. Es war also angezeigt, ein nicht-parametrisches Verfahren zu wählen. Die ergänzende Modellierung mit einem Linearen Modell war nur mit einem eingeschränkten Datensatz bei Imputation fehlender Werte möglich. Aber eben nichtsdestoweniger hilfreich, um sowohl das Teilnahmeverhalten wiederum besser verstehen zu helfen, sowie auch um die Güte des Baum-Modells zu überprüfen.

Anwendung PSA

Mit Hilfe dieser Methoden konnten die Teilnahmewahrscheinlichkeiten zur Partizipation am Panel geschätzt werden. In Kapitel 9 kann dann aufbauend das *propensity score adjustment* angewendet werden. Dazu werden aus den geschätzten Teilnahmewahrscheinlichkeiten Gewichte abgeleitet und bei den Panelisten angewendet. Es ist so möglich, gewichtetet Schätzungen für die Verteilung der Kontrollvariablen vorzunehmen und diese mit dem tatsächlichen Wert, d. h. der Verteilung der Kontrollvariable bei allen CATI-Befragten zu vergleichen. Es ist lässt sich so überprüfen, ob die Gewichtung mittels Teilnahmewahrscheinlichkeiten empirisch zu einer erfolgreichen Reduktion des Bias führen kann. Die Verteilung der Teilnahmewahrscheinlichkeiten selbst, ist nicht nur geeignet, Biasreduzierende Verfahren zu ermöglichen,

15 *Regression trees* heissen eigentlich vollständig *classification and regression trees* und wurden von Breiman et al. (1984) eingeführt.

sondern ist ihrerseits auch ein Indikator für die Repräsentativität einer Stichprobe. Auch solche sogenannte R-Indikatoren können jetzt bestimmt werden, was auch in diesem Kapitel versucht wird.

Simulationen

Überraschenderweise (oder aus Sicht dieser Arbeit vielleicht besser «leider») ist es so, dass es kaum Unterschiede zwischen den CATI-Befragten und den den Panelisten bezüglich der Kontrollvariablen und auch der Kovariaten gibt. Um die Methode des PSA aber noch etwas besser verstehen zu lernen, wurden im Kapitel 10 Simulationen angesetzt. Dazu mussten verschiedene Dinge, die sich empirisch nicht gezeigt haben, gezielt modifiziert werden. Es musste ein stärkerer Zusammenhang zwischen den Kovariaten und der Teilnahmewahrscheinlichkeit konstruiert werden. Ausserdem musste der Unterschied in den Kontrollvariablen zwischen den Panelisten und den CATI-Befragten deutlich verstärkt werden. Dies wurde aufbauend auf die Daten der experimentellen Befragung in verschiedenen Varianten getan. Ausserdem wurde das PSA mit einer anderen, in der Praxis häufig angewendeten Methode der Biasreduktion verglichen (*generalized regression estimator GREG)*, um doch geeignete Hinweise zur Anwendung des PSA in der Praxis der Umfrageforschung geben zu können.

Die Arbeit beginnt im nächsten Kapitel 2 mit einer Einführung in das Problem des Bias. Es bespricht die verschiedenen möglichen Ursachen und Auswirkungen von Bias, speziell auch im Kontext von Web-Panel-Befragungen.

2 Bias

► Bevor die Methode des *propensity score adjustments* eingeführt wird, soll erst das zu lösende Problem diskutiert werden: Bias. Bias wird zunächst allgemein eingeführt und dann im Kontext von Web-Panel-Befragungen besprochen. Im Anschluss werden überblicksartig Methoden zur Reduzierung von Bias vorgestellt, was ins nächste Kapitel überleitet, in dem es um die Schätzung der Teilnahmeneigung gehen wird. ◄

Erwartungstreue

Die Realisierung einer Stichprobe kann man sich als einmaliges Zufallsexperiment vorstellen: einmalig werden aus einer Grundgesamtheit (also dem Teil der Bevölkerung, über die eine Aussage getroffen werden soll) zufällig Elemente (zu befragende Personen) ausgewählt. Man kann dann Eigenschaften dieser ausgewählten Personen messen, sie so z. B. zu ihrem Einkommen befragen und dann Schlussfolgerungen bezüglich der zu messenden Eigenschaft in der Grundgesamtheit treffen. Wenn man (theoretisch) sehr viele solcher Stichproben zieht und in jeder dieser Stichproben diese Eigenschaft messen würde, sollte der durchschnittliche Wert (d. h der Erwartungswert) dieser Eigenschaft über alle theoretischen Stichproben dem tatsächlichen Wert dieser Eigenschaft in der Grundgesamtheit entsprechen. Ist dem nicht so, liegt also eine systematische Verzerrung in den Stichproben vor, spricht man von einem *Bias*.

Bias

Ist es beispielsweise schwieriger, Besserverdienende zu einer Befragung zu motivieren, ist es wahrscheinlich, dass Besserverdienende in der Befragung unterrepräsentiert sind; man schätzt das Durchschnittseinkommen der Bevölkerung dann zu niedrig – jedenfalls dann, wenn man nicht andere externe Informationen zur Korrektur heranziehen kann.

Bias bedeutet[1], dass ein Schätzer, wie zum Beispiel der Schätzer $\hat{t}$ für das Total T, verzerrt ist.

1 Siehe Cochran (1977), Särndal und Lundström (2005) und Lohr (1999)

Erwartungstreue Schätzer

Schätzer, die nicht durch einen Bias beeinflusst werden, können nur bestimmt werden, wenn die Einschlusswahrscheinlichkeit für jedes Element der Grundgesamtheit bekannt und grösser Null ist und wenn ein Zufallsauswahlverfahren zur Ziehung der Stichprobe verwendet wurde (Horvitz und Thompson, 1952). Im Falle von *under-coverage*, d. h. bei den meisten Web-Befragungen, ist dies nicht der Fall[2].

Genau wie bei allen anderen Befragungsmodi auch kommt es bei Web-Befragungen meistens zu *Nonresponse*. Nonresponse bezeichnet das Problem, dass es (meist systematische) Ausfälle von zu befragenden Personen gibt. Da es sich bei Web-Befragungen um eine selbstverwaltete (*self-administered*) Befragungen handelt, ist die Wahrscheinlichkeit von Nonresponse hoch (Bethlehem, 2009).

Probleme durch Nonresponse

Eine spezifische Quelle von Nonresponse bei Web-Befragungen ist, dass manche Befragte Probleme im Umgang mit dem Internet haben. Neben Hindernissen, die aus mangelnder Kompetenz resultieren, gibt es eine Reihe möglicher technischer Probleme. So kann die Übertragungsgeschwindigkeit niedrig sein oder die Darstellung des Fragebogens scheitert an veralteten Browsern, zu niedriger Auflösung usw. Für eine Darstellung vieler möglicher technischer Probleme bei Online-Befragungen siehe z. B. Fricker und Schonlau (2002) und Heerwegh und Loosveldt (2002).

In der Praxis ergeben sich insbesondere auch aus möglichen technischen Schwierigkeiten Probleme bei der Beurteilung der Güte der Befragung. So ist es beispielsweise für den Untersuchungsleiter nicht möglich zu erkennen, ob eine per E-Mail eingeladene Person nicht geantwortet hat, weil sie bewusst verweigert oder weil die Einladungs-E-Mail als Spam deklariert wurde[3] oder ein anderes technisches Problem vorgelegen hat. Bei anderen Befragungsmodi sind solche Komplikationen m. E. besser zu erkennen.

2 Web-Befragungen, bei denen die Einladungen über Bannerwerbung erfolgt oder sich an Personen richtet, die zufällig eine Seite besuchen, sind die Einschlusswahrscheinlichkeiten der einzelnen Elemente unbekannt. Solche Befragungen werden auch *self-selecting* genannt.

3 Viele Befragungen gerade im akademischen Bereich werden mit Befragungstools wie UniPark (`www.globalpark.de`) durchgeführt. Diese Programme verwalten auch die Kommunikation mit den Befragten, d. h. auch die Einladungs-E-Mails werden aus solchen Programmen heraus versendet. Viele Spam-Filter kontrollieren, ob die Domain der angegebenen E-Mail-Adresse der Name des Postausgangsservers ist. Bei derartigen Befragungstools ist dies nicht der Fall, weswegen die Einladungs-E-Mails ohne Hinweis an den Absender gelöscht werden. Es ist dann nicht klar, wie viele der eingeladenen Befragten tatsächlich auch die Einladung jemals tatsächlich gesehen haben.

2.1 Bias als Folge von Nonresponse

Rubin (1987) unterscheidet prinzipiell drei Mechanismen, die zu fehlenden Daten führen können. (Da sich die englischen Bezeichnungen eingebürgert haben, sollen sie hier beibehalten werden.) Daten können *missing completely at random (MCAR)* sein. Bezogen auf eine zu untersuchende Variable Y gilt, dass die Y bestimmenden Variablen X unabhängig sind von der Neigung ρ, an einer Befragung teilzunehmen[4]. MCAR verursacht keinen Bias von Y, ist aber in der Praxis sehr selten. Ein triviales Beispiel für eine solche Unabhängigkeit liegt vor, wenn $\forall_i Y_i \equiv \bar{Y}$, also Y für alle Elemente in der Grundgesamtheit konstant ist.) Liegt MCAR vor, ist die Kovarianz $\sigma_{Y\rho} = 0$. MCAR

Missing at random (MAR) sind Daten dann, wenn sowohl die Ausfallmechanismen (Neigung teilzunehmen) wie auch die zu untersuchende Variable Y konditional von Variablen eines Vektors Z abhängen. Die Bezeichnung «missing at random» ist damit eher irreführend. In der Praxis ist das der gewöhnliche Fall. Im Fall von MAR gilt unter Berücksichtigung von Z $\sigma_{Y\rho} = 0$, ohne Berücksichtigung von Z allerdings $\sigma_{Y\rho} \neq 0$. MAR

Die Neigung an einer Befragung teilzunehmen ρ kann auch abhängig sein von der zu untersuchenden Variable Y. Diese Art von Ausfallmechanismus wird *not missing at random (NMAR)* genannt, der Nonresponse ist dann *nonignorable*. Nur wenn nicht NMAR vorliegt, können Verfahren erfolgreich eingesetzt werden, fehlende Werte zu ersetzen. Es gilt dann immer $\sigma_{Y\rho} \neq 0$. Verfahren, fehlende Werte zu ersetzen, können natürlich immer eingesetzt werden. Liegt NMAR vor, erhöhen sie aber bestenfalls die Varianz resultierender Schätzer. In der Praxis ist wohl immer mit einer Mischform aus allen drei Ausfallmechanismen zu rechnen, es wird also praktisch immer einen nicht korrigierbaren Bias geben. NMAR

Der Bias kann durch die Teilnahmerate beeinflusst werden. Ist sie nahe bei bei 1, haben also (fast) alle Befragten, die eingeladen wurden auch geantwortet, ist der Einfluss des Unterschiedes zwischen Teilnehmern und Verweigerern auf den Bias kleiner, als wenn die Teilnahmerate niedrig ist. Allerdings ist die Teilnahmerate kein ausreichendes Kriterium zur Beurteilung des Bias. (Siehe dazu auch noch Abschnitt 3.5.1, ab S. 44.)

4 Zum Konzept der *propensity scores* (so wird die Teilnahmeneigung technisch formalisiert) siehe noch Abs. 2.3, ab S. 18

2.2 Formalisierung des Bias

Es gibt zwei Möglichkeiten, Bias zu formalisieren[5]. Zunächst kann man vereinfachend ein *fixed response* Model unterstellen. Das bedeutet, dass die Population aus zwei Gruppen besteht, die einen sind die Teilnehmer, die anderen die Nonresponenten. Jedes Element der Grundgesamtheit hat einen Indikator R_k mit $R_k = 1$ für alle Teilnehmer und $R_k = 0$ sonst. Die Teilnahmerate kann also geschrieben werden als

$$\frac{\sum R_k}{\sum R_k + \sum 1 - R_k} \tag{2.1}$$

wobei der Nenner die Summe der beiden Strata Teilnehmer und Nonrespondenten bildet. Wird nun aus der Population eine einfache Zufallsstichprobe gezogen, ist y_k nur für die n_R gezogenen Teilnehmer bekannt, wobei $n_R = \sum a_k R_k$ ist und a_k einen Indikator bezeichnet, der 1 ist, falls k in die Stichprobe gezogen wurde und 0 sonst. Der Mittelwert in der realisierten Stichprobe beträgt dann

$$\bar{y}_R = \frac{1}{n_r} \sum a_k R_k Y_k \tag{2.2}$$

Der Erwartungswert von $\bar{y}_R$ beträgt$\bar{y}_R = E(\bar{Y})$. Der Durchschnitt von y kann also nur mit Hilfe des Response-Stratums geschätzt werden. Der Bias B einer solchen Schätzung beträgt dann

$$B(\bar{y}_R) = \bar{Y}_R - \bar{Y} = \frac{N_{NR}}{N} (\bar{Y}_R - \bar{Y}_{NR}) . \tag{2.3}$$

I Bias abhängig von. . .

Der Bias hängt also von zwei Grössen ab: Dem Anteil der Nonrespondenten und der Stärke Unterschieds zwischen den Teilnehmenden und den Nonrespondenten. Es ist daher beispielsweise möglich, dass wenn der Unterschied des Durchschnitts von Y zwischen den Nonrespondenten und den Teilnehmern sehr klein ist, es trotz hoher Verweigerungsrate nicht zu einer biasbehafteten Schätzung von Y kommt. Ist es in der Praxis noch häufig möglich, N_{NR} aus Gleichung 2.3 abzuschätzen, ist über $\bar{Y}_{NR}$ per definitionem

5 Die Ausführungen folgen im Wesentlichen Bethlehem (2009) und Cassel et al. (1983)

nichts bekannt. Es ist dann nicht möglich zu schätzen, wie gross der Bias ist.

Random Response Model

Dieses *fixed response* Model lässt sich leicht erweitern zu einem (realistischeren) *random response* Model. Dabei wird die Zuteilung zu den beiden Strta nicht als ex ante gegeben angenommen, sondern R_k kann als Realisierung einer Zufallsvariable interpretiert werden. Es wird davon ausgegangen, dass jedes Element k der Grundgesamtheit eine Wahrscheinlichkeit ρ_k hat, falls sie gezogen wird, an einer Befragung zu partizipieren. Der Erwartungswert von $\bar{y}_R$ ist dann

$$E(\bar{y}_R) \approx \tilde{Y} = \frac{1}{N} \sum \frac{\rho_k}{\bar{\rho}} Y_k \tag{2.4}$$

$\bar{\rho}$ ist der Durchschnitt aller Teilnahmewahrscheinlichkeiten in der Population, also $\bar{\rho} = \frac{1}{N} \sum \rho_k$. Der Bias beträgt dann

$$B(\bar{\rho}) = \tilde{Y} - \bar{Y} = \frac{S_{\rho Y}}{\bar{\rho}} = \frac{R_{\rho Y} \sigma_\rho \sigma_Y}{\bar{\rho}}, \tag{2.5}$$

wobei $S_{\rho Y}$ die Kovarianz zwischen ρ und Y ist und $R_{\rho Y}$ der zugehörige Korrelationskoeffizient. σ ist die Standardabweichung.

Der Bias wird also insbesondere durch zwei Dinge bestimmt:

II Bias abhängig von. . .

1. Je höher die Wahrscheinlichkeit ist, dass Personen an der Befragung teilnehmen, je höher also die durchschnittliche Teilnahmeneigung ist, umso kleiner ist der Bias. Ist die Teilnahmewahrscheinlichkeit im Idealfall für alle Personen gleich, existiert kein Bias.

2. Je höher der Zusammenhang zwischen der Teilnahmewahrscheinlichkeit und der zu untersuchenden Variable ist, umso höher ist der Bias. Es existiert also weiterhin kein Bias, wenn die Teilnahmewahrscheinlichkeit und interessierende Variable voneinander unabhängig sind, also $\rho \perp Y$ gilt. Der Bias kann also bei verschiedenen Variablen einer Untersuchung unterschiedlich stark sein.

Grundsätzlich gibt es noch ein drittes Szenario, in dem kein Bias existiert, nämlich wenn Y konstant ist. Dieser in der Realität sehr unwahrscheinliche Fall ist auch deswegen hier uninteressant, da dann keine Stichprobe mehr gezogen werden müsste. Eine Beobachtung würde genügen, die Population zu beschreiben.

Nonresponse führt zwar nicht sicher, aber doch mit grosser Wahrscheinlichkeit zu einem Bias. Ein häufig genutzter Indikator, um die Stärke des Bias abschätzen zu können, ist die Angabe der Teilnahmerate. Das ist bei Panels aus zwei Gründen problematisch: erstens ist die Berechnung der Teilnahmerate willkürlich und zweitens ist nicht sicher, inwiefern Nonresponse tatsächlich zu einem Bias führt.

2.3 Bias bei Web-Panel-Befragungen

Panelisten eines Panels, das auf einer Zufallsstichprobe beruht, werden typischerweise im Rahmen von Befragungen zu anderen Themen rekrutiert (Callegaro und Disogra, 2008). Das bedeutet, dass die Frage, ob eine Person bereit ist an einem Panel teilzunehmen, an eine Befragung zu einem beliebigen Thema angehängt wird. Eine erste, hier als unproblematisch angenommene Auswahl von potenziellen Panelisten, besteht in einer Zufallsstichprobe. In der Marktforschung wird es sich dabei typischerweise um eine RDD-artige Stichprobe handeln, in der akademischen und amtlichen Sozialforschung sind darüber hinaus auch *random-route* und andere vergleichbar aufwändige Auswahlverfahren möglich (Diekmann, 2007).

Bias bei Rekrutierung

Ein erster Bias wird dadurch erzeugt, dass es schon in diesem ersten Schritt zu Nonresponse und gegebenenfalls *under-coverage* kommt. Panelist wird also nur jemand, der auch breit ist, an Telefonbefragungen (oder in welche Art von Medium diese Erstbefragung auch sonst eingebettet ist) teilzunehmen und erreicht werden kann. Jemand, der bereit wäre an einer Internet-Befragung teilzunehmen, jedoch das jeweilige Medium ablehnt, hat keine Chance in das Panel zu gelangen. Dillman (2000) und Heerwegh (2009) zeigen (neben anderen), dass sich die Bereitschaft an Interviews teilzunehmen bei *self-administered* Interviews (zu denen Web-Befragungen gehören) und solchen mit Interviewern (z. B. Telefonbefragungen) unterscheiden kann.

Bias durch Panelverweigerung

Ein weiterer Bias entsteht dadurch, dass nicht alle Befragte bereit sind, am Online-Panel teilzunehmen. Selbst wenn das befragende Institut Personen ohne Internetanschluss einen privaten Computer zur Verfügung stellt und einen Internetanschluss bezahlt, sind Befragte ohne eigenen Internet-

anschluss im Panel unterrepräsentiert (Scherpenzeel, 2008)[6]

In der Praxis ist der Ausfallprozess ein doppelter: Zunächst lehnen einige Befragte die Teilnahme am Panel von Anfang an ab. Ein weiterer Teil derjenigen, die die Teilnahme zugesagt haben, sind dann aber doch nicht bereit, an einer Befragung teilzunehmen. LINK berichtet aus Erfahrung, dass rund 40% derjenigen, die ursprünglich zugesagt haben, bei einem ersten Kontakt per E-Mail nicht antworten.

Ein weiterer Bias entsteht auf dieser Stufe dadurch, dass häufig nicht alle Befragte als Panelisten in Frage kommen. Insbesondere wird bei Marktforschungsinstituten eine Mindestnutzung des Internet gefordert oder es werden bestimmte Berufsgruppen wie Journalisten, Marktforscher, etc. ausgeschlossen.

Bias durch Panelabrieb

Besteht das Web-Panel einmal, entsteht wie bei jedem Panel eine Panelmortalität (Sobol, 1959). Im Falle eines Panels können das alle Gründe sein, die auch bei einem anderen Panel in Frage kommen, wie Verweigerung, Tod, Erkrankung usw. Zusätzlich kann es zu technischen Schwierigkeiten kommen, wie dem Wechsel der E-Mail-Adresse. Wenn eine Person nicht auf eine Einladung zu einer Befragung reagiert, ist nicht ohne weiteres klar, ob es sich um einmaligen Nonresponse handelt oder ob die Person dauerhaft aus dem Panel ausgefallen ist. Panelschwund (*panel attrition*) ist ein mindestens kurzfristig nicht zu quantifizierendes Problem.

Bias bei Befragung aus Panel

Aus dem Panel werden dann Befragte mittels einer weiteren Zufallsauswahl ausgewählt und zu der eigentlichen Befragung eingeladen. Auch dabei gibt es wieder Bias verursachenden Nonresponse. In vier Stufen sind also Entscheidungen von zu befragenden Personen einbezogen, die Bias erzeugen können.

Die Unterstellung eines starken Bias bei Web-Befragungen erfolgt häufig nur aus theoretischen Überlegungen (Rhall und Fine, 2008; Scherpenzeel, 2008). Explizite Studien zu den Unterschieden zwischen Antwortenden und Verweigerern und Personen, die nicht erreicht wurden, existieren nur sehr

6 Das mit 13 Mio. Euro geförderte Projekt **CentERdata** versucht ein repräsentatives Web-Panel aufzubauen. Die potentiellen Panelisten werden mittels random-Walk kontaktiert. Allen, die keinen eigenen Computer haben, wird ein solcher kostenlos bereitgestellt. Die Computer wurden extra so entwickelt, dass die Handhabung besonders einfach ist, um auch Personen zu erreichen, die den Umgang mit Computern nicht gewöhnt sind. Nach dem Start stehen z. B. nur ein rudimentäres Schreibprogramm und der Zugang zum Internet zur Verfügung (Toepoel et al., 2009).

wenige. Häufig beziehen sich entsprechende Studien auf die Untersuchung der Unterschiede bei Mehr-Modus Befragungen (Fricker et al., 2005; Klein et al., 2004; Yun und Trumbo, 2000). Dabei ist es aber schwierig abzuschätzen, inwiefern der Unterschied in Antworten bei den unterschiedlichen Befragungsmodi durch einen unterschiedlichen Nonresponse oder durch Unterschiede im medienbedingten Antwortverhalten resultiert.

Couper (2007), Lee (2006a) und Hoogendoorn und Daalmans (2009) untersuchen ähnliche Variablen, um Unterschiede zwischen Respondenten und nicht-Respondenten zu finden. Alle kommen zu ähnlichen Ergebnissen (die auch durch eigene Erfahrung gestützt werden kann):

- Es ist sehr schwierig, ältere Befragte (insbesondere der Gruppe 65+) via Web-Befragungen zu erreichen.
- Es ist problematisch, Personen mit hohem Einkommen bei Web-Befragungen zu erreichen, noch schwieriger aber Personen mit sehr niedrigem Einkommen.
- Es gibt in allen untersuchten Ländern regionale Unterschiede. Innerhalb aller Länder ist es unabhängig von der Region so, dass Befragte aus ländlichen Gebieten etwas leichter zu erreichen sind.
- Personen mit ausländischer Nationalität sind in allen Ländern schwer zu erreichen.
- Junge Männer sind schwer für Web-Befragungen zu gewinnen.

Bias v. Teilnahmerate

Groves (2006) und Groves und Peytcheva (2008) zeigen in zwei aufbauenden Metastudien, dass die Teilnahmeraten den Bias nur schlecht vorhersagen können. Letztere zeigen anhand von 59 Studien, die den Zusammenhang zwischen Nonresponse und Bias untersuchen, dass der Unterschied zwischen denjenigen, die sich an der Befragung beteiligt haben und denjenigen, die sich nicht an der Befragung beteiligt haben, unabhängig vom Anteil der Nichtantwortenden ist.

Incentives

Eine (verblüffende) Ausnahme bildet die Verwendung von materiellen Anreizen (*incentives* bei den Befragungen. Es gilt zwar, dass umso höher die *incentives* sind, desto höher wird die Teilnahmerate. Es scheint aber auch so zu sein, dass je höher die *incentives* sind, desto eher fühlen sich Personen aus den unteren sozialen Schichten angesprochen und umso höher wird ihr

Anteil an den Befragten. Mit steigender Teilnahmerate kann demnach auch der Bias steigen.

Beschreibung der Teilnahmewahrscheinlichkeit

Abbildung 2.1 zeigt nochmals die einzelnen Stufen der Rekrutierung, bei denen es zu einem Ausfall von Befragten kommen kann. Jedes Element i der Grundgesamtheit hat eine Wahrscheinlichkeit ρ_i^1, an einer Befragung teilzunehmen. $\rho^{1.1}$ ist die Wahrscheinlichkeit, zu dieser Befragung eingeladen zu werden und beinhaltet auch mögliche Probleme des *under-* bzw. *over-coverage*. Wird i ausgewählt, hat es eine Wahrscheinlichkeit $\rho_i^{1.2}$, auf die Einladung zu reagieren und tatsächlich an der Befragung teilzunehmen. $\rho^{1.2}$ bildet auch die Wahrscheinlichkeit ab, vom Interviewer erreicht zu werden und in der Lage zu sein teilzunehmen. Bei Web-Panel-Befragungen handelt es sich bei $\rho^1 = \rho^{1.1}\rho^{1.2}$ um die Wahrscheinlichkeit, an einem Rekrutierungsinterview teilgenommen zu haben.

Hat i teilgenommen, stimmt es mit der Wahrscheinlichkeit $\rho_i^{2.1}$ zu, an einem Panel teilzunehmen. Mit der Wahrscheinlichkeit $\rho_i^{2.2}$ verbleibt i bis zum Zeitpunkt der eigentlichen Web-Befragung im Panel. Ist dies der Fall und wird i aus dem Panel ausgewählt, an einer Befragung teilzunehmen, reagiert i mit der Wahrscheinlichkeit $\rho_i^{2.3}$ positiv. (Die Auswahl von i aus dem Panel, an der Befragung teilzunehmen, wird als nicht biasbehaftet betrachtet, da in einem gut gepflegten Panel eine Liste aller Panelisten vorhanden und damit eine Zufallsauswahl möglich ist.)

Die absolute Teilnahmewahrscheinlichkeit ρ ist das Produkt all dieser Wahrscheinlichkeiten. Der Einfachheit halber sei angenommen, dass die einzelnen Wahrscheinlichkeiten unabhängig voneinander sind.

$$\rho_i = \rho_i^{1.1} \times \rho_i^{1.2} \times \rho_i^{2.1} \times \rho_i^{2.2} \times \rho_i^{2.3} \tag{2.6}$$

In Anlehnung an die englische Bezeichnung *propensity score* soll diese Wahrscheinlichkeit auch «Teilnahmeneigung» genannt werden.

2.4 Methoden zur Reduktion von Bias

Typischerweise werden den Befragten Gewichte zugeschrieben, um den durch Nonresponse und *under-coverage* hervorgerufenen Bias zu beseitigen. Die Gewichte werden so entwickelt, dass die Verteilung definierter Merkmale

in der Stichprobe der bekannten Verteilung in der Population entspricht.

Horvitz-Thompson Schätzer

Wenn in der Stichprobe mit keinerlei Verzerrung zu rechnen ist, ist der Horvitz-Thompson Schätzer ein Schätzer ohne Bias (Horvitz und Thompson, 1952). Er verwendet die (bekannten) Einschlusswahrscheinlichkeit π_i eines Elements, um Gewichte zu bestimmen. Der Horvitz-Thompson Schätzer für das Populationsmittel ist

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{\pi_i} \tag{2.7}$$

Der Horvitz-Thompson Schätzer ist erwartungstreu für das Populationsmittel. Die Gewichte $w_i = 1/\pi_i$ werden Hochrechnungsgewichte genannt (Hulliger, 2006).

GREG

Um einen möglichen Bias korrigieren zu können, führen Särndal (1980) und Bethlehem (1988) den *Generalized Linear Regression Estimator (GREG)* ein. Der GREG Schätzer entspricht dem Horvitz-Thompson Schätzer, ergänzt um ein Korrekturterm, der auf externen Informationen beruht. Der Schätzer für das Populationsmittel erweitert sich zu:

$$\bar{y}_{gr} = \bar{y}_{ht} + \left(\bar{X} - \bar{x}_{ht}\right)^T \beta \tag{2.8}$$

β muss mit Hilfe der Stichprobe geschätzt werden als

$$\hat{\beta}^S = \left(\sum_{k=1}^{N} \frac{a_k X_k X_k^T}{\pi_k} \right)^{-1} \left(\sum_{k=1}^{N} \frac{a_k X_k Y_k}{\pi_k} \right) \tag{2.9}$$

Der GREG Schätzer benötigt externe Informationen, die mindestens als Total oder Mittelwert für die Population bekannt sein müssen. Wenn ein Unterschied zwischen den mit Hilfe der Stichprobe geschätzten Mittelwerte

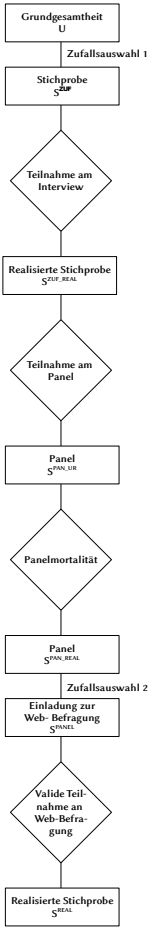


Abbildung 2.1: Bias bei Web-Panel Befragungen

und den tatsächlichen Mittelwerten der Hilfsvariablen besteht, kann der Horvitz-Thompson Schätzer entsprechend korrigiert werden. Korrelieren die Hilfsvariablen und die interessierende Variable, kann man davon ausgehen, dass auch sie einem ähnlichen Bias wie die Hilfsvariable unterliegt und entsprechend korrigiert wird. Für weitere Details siehe zum Beispiel Cobben (2009) oder Bethlehem (2009).

Der GREG Schätzer ist wie zum Beispiel das *iterative proportional fitting (IPF)* oder die Poststratifizierung ein Spezialfall des *calibration framework*, eingeführt von Deville und Särndal (1992). Sie erweitern den Horvitz-Thompson Schätzer (Gl. 2.7) zu

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} w_i Y_i \tag{2.10}$$

wobei sich die Gewichte w_i zusammensetzen aus $w_i = \frac{1}{\pi_i} g_i$. Zu dem Horvitz-Thompson Hochrechnungsgewicht wird ein zweites Korrekturgewicht ergänzt (*g-weight*). Die Gewichte müssen folgende Kallibrierungs-Gleichung erfüllen:

$$\sum_{i=1}^{n} w_i \mathbf{X}_i^n = \sum_{i=1}^{N} \mathbf{X}_i \tag{2.11}$$

Die Totale der gewichteten Hilfsvariablen in der Stichprobe sollen also den Totalen der Hilfsvariablen in der Population entsprechen. Die Totale der Population müssen daher bekannt sein. Es gibt meist nicht nur eine mögliche Lösung für die Gleichung, d. h. verschiedene Verteilungen von Gewichten können zu diesem Ergebnis führen. Die meisten Methoden unterstellen daher noch eine Art von Distanzfinktion zwischen w_i und $1/\pi_i$, die minimiert werden muss.

Weitere Alternativen

Weitere Methoden, um den Bias zu reduzieren, werden z. B. von Bethlehem (2002), Bethlehem (2007) und Särndal und Lundström (2005) vorgestellt. Die meisten unterscheiden sich von dem gleich zu beschreibenden *propensity score adjustment* dadurch, dass weniger Informationen benötigt werden. Typischerweise werden nur Totale bzw. Mittelwerte der externen Variablen benötigt. *propensity score adjustment* benötigt dagegen Informationen über die Verteilung der Kovariate auf Individualebene. Die $\mathbf{X}_i$ müssen für alle $i \in U$ bekannt sein oder mindestens für eine repräsentative Teilmenge von U.

3 Response Propensity -- Teilnahmeneigung

► Ein Ziel dieser Arbeit ist es zu zeigen, dass mit Hilfe der Schätzung der Wahrscheinlichkeit, an einer Befragung teilzunehmen, eine Biasreduktion vorgenommen werden kann. Die dies ermöglichende Methode heisst *propensity score adjustment* und wird in diesem Kapitel vorgestellt werden. Es werden die Voraussetzungen benannt, die einer Schätzung der Teilnahmeneigung zu Grunde liegen. Es werden dann statistische Methoden vorgestellt, mit deren Hilfe die Teilnahmeneigung geschätzt werden kann. Der letzte Abschnitt dieses Kapitels behandelt sogenannte R-Indikatoren. R-Indikatoren sind ein Mittel, die Repräsentativität von Befragungen zu schätzen. Auch sie basieren auf der Schätzung der Teilnahmewahrscheinlichkeit. ◄

Die Teilnahmewahrscheinlichkeit wird dabei als *propensity scores* abgebildet, siehe auch Abschnitt 2.3, ab S. 18.

Referenz

Das Konzept der *propensity scores* wurde als erstes von Rosenbaum und Rubin (1983) entwickelt, siehe auch Rosenbaum und Rubin (1984) und D'Agostino und Rubin (2000). Das Ziel war es, bei experimentellen Studiendesigns mit evaluatorischem Charakter Bias zu reduzieren, wenn es nicht oder nur eingeschränkt möglich ist, die Auswahl der Probanden zufällig auf die Untersuchs- und die Kontrollgruppe zu verteilen. Die bei evaluatorischen Experimenten interessierende Statistik ist die Wirkung der zu untersuchenden Massnahme. Damit dabei kein Bias die Ergebnisse beeinflusst, wurden *propensity scores* eingeführt. Die ersten, die das Konzept auf Befragungen übertragen haben, um Probleme durch Nonresponse zu beseitigen, waren David et al. (1983).

Propensity scores wurden schon wiederholt zur Bestimmung von Gewichten bei Web-Befragungen verwendet (Schonlau et al., 2004; Oberski, 2006; Terhanian, 2000). Allerdings sind noch viele Probleme ungelöst und weitere Studien nötig (Lee, 2006b). Insbesondere ist nicht klar, mit Hilfe welcher

Variablen die *propensity scores* bestimmt werden sollen. Entweder handelt es sich um ausschliesslich demografische Variablen wie bei Oberski (2006) oder das Verfahren ist im Rahmen privater Marktforschung eingesetzt worden und daher nur schlecht dokumentiert, wie es z. B. bei den Pionieren der Anwendung des Verfahrens, *Harris Interactive*, der Fall ist.

3.1 Propensity Scores als Mittel der Bias-Reduktion

Formalisierung

Mit einer Wahrscheinlichkeit ρ_i nimmt eine ausgewählte Person i an der Befragung teil. Diese Teilnahmewahrscheinlichkeit kann als bedingte Wahrscheinlichkeit modelliert werden. Man nimmt an, dass ρ_i vollständig durch die Bedingung X_i definiert ist (Cobben, 2009; Bethlehem, 2007; Rosenbaum und Rubin, 1983).

$$\rho(X_i) = Pr\,(I_i = 1|X_i) \tag{3.1}$$

I ist eine Indikatorvariable, die den Wert 1 annimmt, falls i ein antwortet und anderenfalls 0. Zur Auswahl der Kovariate siehe Kapitel 4, ab S. 55.

ρ ist eine latente Variable, also nicht unmittelbar beobachtbar. Es ist aber möglich, ρ als $\hat{\rho}$ mittels bekannter Kovariate $\mathbf{X}$ zu schätzen.

$$\hat{\rho}_i = Pr\,(I_i = 1|\mathbf{X}_i)\,, \tag{3.2}$$

Befragte mit der gleichen Kovariatenverteilung $\mathbf{X}_i$ sollten die gleiche Wahrscheinlichkeit haben, an der Befragung zu partizipieren. Das ist die *missing at random (MAR)* Annahme aus Abschnitt 2.1. Zur Schätzung von *propensity scores* siehe auch noch ausführlicher die Kapitel 7 (ab S. 119), und 8 (ab S. 137).

Basisannahme

Tatsächlich ist es so, dass die Teilnahmewahrscheinlichkeit nicht nur von beobachtbaren Kovariaten X abhängt, sondern darüber hinaus von nicht beobachtbaren Eigenschaften Z und meistens auch vom Thema der Befragung, also Y. *Propensity score adjustment* kann nur angewendet werden, wenn $\rho(X,Y,Z)$ angenähert werden kann durch $\rho(X)$. Dass man vollständig alle Kovariaten erfassen kann, die das Teilnahmeverhalten bestimmen, ist in der Praxis illusorisch. Daher können Verfahren, die den Bias auf Basis der

Bestimmung der Teilnahmeneigung reduzieren, nur zu einer Verbesserung des Bias führen, diesen aber nicht vollständig eliminieren.

Eine Stichprobe kann einen unterschiedlichen Grad an Reräsentativität für die verschiedenen erhobenen Variablen haben. Sie kann gelegentlich die eine Variable sehr gut abbilden, z. B. eine, die völlig unabhängig von der Teilnahmewahrscheinlichkeit ist (beispielsweise der Blutgruppe), und gleichzeitig für eine andere Variable zu vollständig falschen Schlüssen verleiten, wie es z. B. meist bei Einstellungsfragen der Fall ist oder auch trivialerweise bei der Frage «Haben Sie schon jemals an einer Befragung teilgenommen?». Idealerweise wäre ρ daher definiert als $\rho(y_i) = E(I_i|y_i)$. Die Teilnahmewahrscheinlichkeit wäre also definiert in Abhängigkeit der interessierenden Variable y. Bestenfalls müsste die Teilnahmewahrscheinlichkeit also sogar für jede interessierende Variable y gesondert bestimmt werden. Um ρ so bestimmen zu können, müsste y_i für alle Elemente der Population bekannt sein, was annahmegemäss nie der Fall sein wird – es wäre ja sonst keine Befragung mehr nötig. Aus diesem Grund muss ρ über den Umweg Kovariate geschätzt werden.

PS als Mittel der Biasreduktion

Es gibt mehrere Möglichkeiten, den *propensity score* für eine Bias-Reduktion zu verwenden. Es sollen im Folgenden verschiedene Varianten kurz besprochen werden

Gewichtung Bei dieser Methode handelt es sich um das Porpensity Score Weighting im engeren Sinne (Bethlehem, 2009). Die aus den *propensity scores* abgeleiteten Gewichte berechnen sich einfach als $w_i^{PS} = \frac{1}{\hat{\rho}}$. Der typischerweise verwendete Horvitz-Thompson Schätzer zur Schätzung des Mittelwerts einer Population ändert sich dann in

$$\overline{y}_{ht} = \frac{1}{N} \sum_{i \in S_r} \frac{a_i Y_i I_i}{w_i \hat{\rho}_i} \tag{3.3}$$

a_i bezeichnet einen weitere Indikatorvariable, die 1 beträgt, falls i für die Stichprobe ausgewählt wurde und anderenfalls 0 ist.

Stratifizierung Die Gewichtung mittels *propensity scores* wird häufig so vorgenommen, dass die Gewichte nicht unmittelbar aus den Scores abgeleitet werden, sondern dass die berechneten Scores in H Gruppen eingeteilt werden und eine Poststratifizierung vorgenommen wird. Man spricht auch von *propensity score stratification*. Die Strata müssen so

konstruiert werden, dass die Variablität der Teilnahmewahrscheinlichkeiten innerhalb eines Stratums klein und zwischen den Starta gross ist. Cochran (1968) zeigt, dass fünf Strata bei Poststratifizierungen ausreichend sind, um 90 % des Bias zu reduzieren. Cobben und Bethlehem (2005) zeigen, dass 25 Strata zu einem etwas besseren Resultat führen.

Jedes Element einer Schicht h erhält identische Gewicht

$$g_h = \frac{N_h}{N_{h,r}} \tag{3.4}$$

$n_{h,r}$ gibt dabei die Grösse der Schicht im Sample und n_h in der Grundgesamtheit an. Der *response propensity stratification estimator* (Cobben, 2009) ergibt sich dann als

$$\overline{y}_{PS}^{\hat{\rho}} = \frac{1}{N} \sum_{h=1}^{H} n_{h,r} g_h \overline{y}_r^h \tag{3.5}$$

wobei $\overline{y}_r^h$ das ungewichtete Mittel von y in der Schicht h ist.

Für den Fall der Poststratifizierung zeigt Bethlehem (1988), dass wenn m die Grösse der Referenzbefragung bezeichnet, die Varianz V des resultierenden Schätzers für den Mittelwert berechnet werden kann als

$$V\left(\overline{y}_{PS}^{\hat{\rho}}\right) = \frac{1}{m} \sum_{h=1}^{H} g_h \left(\overline{y}_h - \overline{y}\right)^2 + \frac{1}{m} \sum_{h=1}^{H} \left(1 - g_h\right) V\left(\overline{y}_h\right) + \frac{1}{m} \sum_{h=1}^{L} g_h^2 V\left(\overline{y}_h\right) \tag{3.6}$$

Der erste Term der Gleichung wird umso kleiner, je grösser m ist. Umso grösser also die Referenzbefragung ist, umso kleiner ist die Varianz. Der dritte Term der Gleichung enthält $\overline{y}_h$, einen Wert, der in der Stichprobe gemessen wurde. Dessen Varianz $V(\overline{y}_h)$ ist also von der Ordnung $\frac{1}{n}$ und wird also umso kleiner, je grösser die Stichprobe der Web-Befragung ist. Der mittlere Term wird umso kleiner, je grösser sowohl m wie auch n sind.

Um eine möglichst niedrige Varianz des Schätzers zu erreichen, müs-

sen also sowohl die eigentliche Stichprobe wie auch die Referenzbefragung möglichst gross sein.

In einer anderen Untersuchung zeigen Cobben und Bethlehem (2005), dass *propensity score stratification* zu unbefriedigenden Resultaten bezüglich der Biasreduktion kommen kann, wenn nicht weitere Variablen zur Bildung der Starta herangezogen werden.

Lineare Gewichtung I Cobben und Bethlehem (2005) schlagen im Zusammenhang mit Telefonbefragungen vor, eine GREG-Schätzer so zu modifizieren, dass auch *propensity scores* berücksichtigt werden. β wird jetzt nicht mehr wie in Gl. 2.9 nur mit Hilfe der einfachen Einschlusswahrscheinlichkeit π_k gemäss Stichprobenplan geschätzt. Um die Teilnahmewahrscheinlichkeit zu berücksichtigen, wird diese mit der Teilnahmewahrscheinlichkeit ρ_k multipliziert. Da ρ unbekannt ist und als $\hat{\rho}_k$ mit Hilfe von X_k geschätzt werden muss, ergibt sich zur Bestimmung von $\hat{\beta}$ jetzt

$$\hat{\beta} = \left(\sum_{k=1}^{N} \frac{a_k X_k X_k^T}{\pi_k * \hat{\rho}_k} \right)^{-1} \left(\sum_{k=1}^{N} \frac{a_k X_k Y_k}{\pi_k * \hat{\rho}_k} \right) \tag{3.7}$$

Lineare Gewichtung II In einer auch von Cobben und Bethlehem (2005) vorgeschlagenen Variation wird ein GREG Schätzer verwendet, wobei kategorisierte *propensity scores* als stratifizierende Variablen verwendet werden.

3.2 Voraussetzungen

Werden *propensity scores* für eine Gewichtung verwendet, sind sowohl die Kovariaten wie auch die Untersuchungsergebnisse unabhängig von der Teilnahme bzw. Nicht-Teilnahme an der Befragung: $\mathbf{X} \perp I|\rho$ und $Y \perp I|\rho$. Vergleiche dazu auch die Theoreme 1 und 3 in Rosenbaum und Rubin (1983).

Drei Annahmen müssen getroffen werden, damit die Schätzung einer interessierenden Statistik nach einer *propensity scores* Korrektur möglich ist:

1. Es darf keine unbeobachtete Kovariate geben, die die Neigung teilzunehmen beeinflusst.

2. Der Modus der Befragung darf keinen Einfluss auf das Antwortverhalten haben. Sollte es einen solchen Einfluss geben, wäre eine Berechnung der *propensity scores* nicht möglich. Siehe dazu noch die Diskussion in Abschnitt 3.2.2

3. Jedes Mitglied der zu untersuchenden Population muss eine Wahrscheinlichkeit grösser als Null haben, an einer Befragung teilzunehmen.

$$0 < \rho(I_i = 1|\mathbf{X}) \leq 1 \quad \text{für alle} \quad i \in U \tag{3.8}$$

   Sollte es eine Gruppe von Personen geben, die via Befragung nicht zu erreichen ist, kann der dadurch verursachte Bias nicht korrigiert werden.

Diese drei Bedingungen sollen im Folgenden kurz besprochen werden.

3.2.1 Unbeobachtete Kovariate

Eine erste Voraussetzung ist das Vorhandensein externer Informationen.

Bei Web-Befragungen kann die Schätzung der Teilnahmeneigung häufig nur durch eine kombinierte Online- und Offline-Befragung realisiert werden, wobei vorausgesetzt wird, dass die Offline-Befragung valide Daten produziert. Die Biasreduktion mittels *propensity score weighting* kann dann nur bis auf die Ebene der Offline-Befragung vorgenommen werden. Daher müssen die Ansprüche an die Offline-Befragung sehr hoch sein.

Die geschätzte Teilnahmeneigung bildet dann in Analogie zu Gleichung 2.6 nicht mehr die vollständige Teilnahmewahrscheinlichkeit ρ ab, sondern nur einen Teil davon, nämlich

$$\hat{\rho} = \rho_i^{2.1} \times \rho_i^{2.2} \times \rho_i^{2.3} \tag{3.9}$$

Qualität von Offline-Befragungen

Es ist nicht sicher, dass Offline-Befragungen per se von höherer Qualität sind als Web-Befragungen. Zum Beispiel werden Fragen, deren Beantwortung sozialer Erwünschtheit unterliegt, besonders offen online beantwortet

(Tourangeau und Smith, 1998). Die Erreichbarkeit der allgemeinen Bevölkerung ist offline sicherlich besser (z. B. Rhall und Fine (2008) oder Scherpenzeel (2008)), was der Grund für die höhere Akzeptanz solcher Befragungsmodi ist. Und auch der Grund dafür, dass eine Offline-Befragung auch hier als Referenz benutzt wird.

Fricker et al. (2005) haben eine RDD-Befragung durchgeführt, bei der sie als erstes gefragt haben, ob die Befragten auch mit einer Web-Befragung einverstanden sind. Alle die einverstanden waren, wurden zufällig entweder per Telefon oder per E-Mail befragt. Es zeigt sich, dass die Teilnahmebereitschaft bei denen, die telefonisch befragt wurden, deutlich höher ist. (Allerdings war der Item-Nonresponse bei der Web-Befragung niedriger.) Allerdings haben sie auch Hinweise dafür gefunden, dass Wissensfragen gewissenhafter online beantwortet wurden.

3.2.2 Moduseffekte

Die Rekrutierungsbefragung bildet die Grundlage für die spätere Schätzung der Neigung zur Teilnahme an Panels. Die Variablen, die sich als gute Prädiktoren erweisen, werden dann bei jeder Web-Befragung erneut gestellt, um eine Gewichtung (oder Matching) ableiten zu können. Dabei stellt sich die Frage, inwieweit Befragungen via Internet und via Telefon vergleichbar sind, d. h., es stellt sich die Frage, ob es nicht zu Unterschieden im Antwortverhalten nur auf Grund des benutzten Mediums kommen kann, also Moduseffekte vorliegen.

Ungewohnter Befragungsmodus

Eine wichtige Quelle von Vorurteilen gegenüber Web-Befragungen besteht darin, dass es (immer noch) ungewohnt erscheint, mittels Internet zu kommunizieren, oder mindestens, dass dies für einige Gruppen der Bevölkerung zutrifft (de Leeuw, 1992). Gemeinhin werden Befragungen von Angesicht zu Angesicht und via Telefon (Groves, 1989) als «natürlichere» d. h. gewohntere Komminikationsformen betrachtet. Ähnliches gilt für die schriftliche Kommunikation, insbesondere das Ausfüllen von Formularen. Auch das scheint im Alltag mindestens der Bewohner westlicher Länder ein alltäglicher und damit vertrauter Kommunikationsakt zu sein (de Leeuw, 1992). Und aus dieser Vertrautheit mit dem Medium wird eine höhere Reliabilität abgeleitet (de Leeuw, 1992).

Welche Moduseffekte es zwischen Internet- und CATI-Befragungen gibt,

ist nicht hinreichend untersucht. Klar ist zumindest, dass es Unterschiede zwischen persönlichen «face-to-face» Befragungen und selbst verwalteten Befragungen, wie telefonischen und schriftlichen Befragungen, geben kann (Rosenthal, 1966). Ein Beispiel ist der bewusste oder unbewusste Einfluss des Interviewers auf die Befragten
(Downs et al., 1980; Bungard, 1980). Solche «Versuchsleiter-Artefakte» (auch «Rosenthal-Effekte») existieren bei schriftlichen Befragungen nicht und bei telefonischen nur eingeschränkt (Rosenthal, 1966).

Mögliche Vorteile durch Online-Befragungen

Ein Beispiel ist die positive Verstärkung durch räumliche Nähe zwischen Interviewer und Interviewtem und die Beeinflussung der für das Antwortverhalten relevanten Interviewatmosphäre durch die Häufigkeit der Blickkontakte. Rosenthal und Bungard haben lange Listen solcher Versuchsleiter-Artefakten aufgestellt (Rosenthal, 1966; Bungard, 1980). Als Vergleichsbefragung zur internetbasierten Befragung eine persönliche Befragung zu verwenden, empfiehlt sich daher nicht.

Abschliessende befriedigende empirische Tests zu den Moduseffekten von internetbasierten Befragungen im Vergleich zu anderen Befragungsmodi existieren meines Wissens nicht[1]. Die bisher gefundenen Ergebnisse sind nicht einheitlich.

Manfreda und Vehovar (2002a) vergleichen in einer Studie die Ergebnisse einer Online-Befragung mit den Ergebnissen sonst identischer postalischer und telefonischer Befragungen. Bei 11 der 27 gestellten Fragen unterscheiden sich die Antworten online und schriftlich. Gegenüber der telefonischen Befragung unterscheiden sich die online gegebenen Antworten nur bei 8 Fragen. (Leider publizieren Manfreda und Vehovar nicht, inwieweit sich die Antworten zwischen telefonischer und postalischer Befragung unterscheiden.)

Allerdings erfolgte die Einladung zur webbasierten Befragung mittels (Internet-) Bannerwerbung. Bei dieser Art der Selbstrekrutierung werden nahezu ausschliesslich internetaffine Personen, die eine hohe Bereitschaft zur Teilnahme an Befragungen haben, ausgewählt. Effekte, die z. B. dadurch entstehen, das ungewohnte Werkzeug Computer zu benutzen, sind in

1 Eine Sammlung vieler Artikel, Beiträge und Bücher zum Thema Web-Befragungen findet man unter `www.WebSM.org`. Die Seite wird von Vasja Vehovar betrieben und soll eine umfassende Sammlung aller Publikationen zu Web-Befragungen sein. Dort wurden keine weiteren Artikel, ausser den unten aufgeführten, gefunden.

der Studie daher ausgeschlossen. Wie stark der mediumsbedingte Unterschied zwischen den Antworten tatsächlich ist, ist daher fraglich. Ausserdem wurden die nicht internetbasierten Befragungen erst viele Wochen nach der Online-Befragung durchgeführt, ohne nochmals eine Online-Kontroll-Gruppe zu bilden. Es ist daher nicht ausgeschlossen, dass die Unterschiede in den Antworten auch durch andere Dinge beeinflusst wurden.

Empirische Befunde

Vehovar et al. (1999) haben mit einen ähnlichen Untersuchungsdesign internetbasierte, postalische und telefonische Befragungen verglichen. Die Unterschiede zwischen der postalischen und der internetbasierten Befragung sind zwar deutlich, zwischen der telefonischen und der webbasierten Befragung sind die Unterscheide dagegen wieder nur sehr gering.

Speizer et al. (2005) finden auch Unterschiede in den Antworten zwischen einer internetbasierten und einer identischen telefonischen Befragung. Allerdings trennen sie nicht zwischen Unterschieden, die durch den Befragungsmodus entstehen, und solchen, die durch einen onlinebedingten Bias resultieren.

Taylor (2000) findet keine Unterschiede zwischen telefonischen und internetbasierten Befragungen, wenn man mittels demographischer Angaben bestimmte Gewichte berücksichtigt. Genau wie Taylor haben auch Varedian und Forsman (2002) eine internetbasierte Befragung mit einer *random digit dialing (RDD)* Befragung verglichen[2]. Sie kommen zu dem Resultat, dass die verschiedenen Befragungsmodi in einigen Fällen zu unterschiedlichen Ergebnissen führen können.

Bei einer Befragung zum Umweltbewusstsein in Deutschland haben Bandilla et al. (2003) bei 67% der Fragen unterschiedliche Antworten beim Vergleich von einer internetbasierten mit einer postalischen Befragung erhalten. Auch sie haben die Daten mittels Poststratifikation gewichtet.

Schonlau et al. (2002) vergleichen eine internetbasierte Befragung mit einer RDD-Befragung. Die Online-Befragung ist in Zusammenarbeit mit Harris Interactive[3] entstanden, der Marktforschungsfirma, die bereits Gewichtungen auf Basis von *propensity scores* vornimmt, ohne das Vorgehen allerdings öffentlich zu dokumentieren. Auch bei dieser Vergleichs-Studie

2 Die Befragungen sowohl von Taylor als auch von Varedian und Forseman hatten den Gebrauch und Konsum von Hygieneartikel zum Thema.

3 www.harrisinteractive.com

wurde eine Gewichtung nach *propensity scores* vorgenommen, aber es bleibt unklar, welche Variablen als Kovariate herangezogen wurden.

Von den 37 gestellten Fragen haben sich die Antworten bei je nach Berechnung acht oder elf Fragen nicht unterschieden. Bei den Fragen mit Unterschieden handelte es sich ausschliesslich um solche mit fünfstufiger Rating-Skala. Wurden bei diesen Skalen die Kategorien zusammengefasst, also z. B. «very good» und «good», zeigen sich keine signifikanten Unterschiede mehr zwischen den Befragungsmodi.

3.2.3 Positive Teilnahmewahrscheinlichkeit

Die letzte Bedingung scheint auf den ersten Blick am wenigsten realistisch zu sein. In Anbetracht von *under-coverage* und Nonresponse ist es zunächst offensichtlich, dass es Personen gibt, die tatsächlich nie an einer Web-Befragung teilnehmen würden, für die also $\rho = 0$ gilt. Dazu zwei Bemerkungen:

Under-coverage Natürlich gibt es Personen, die das Internet eigentlich nie benutzen. Diese haben dann eine Wahrscheinlichkeit von Null, an einer Befragung teilzunehmen, bei der die Einladung z. B. via Banner oder E-Mail kommuniziert wird. Im Kontext von offline rekrutierten Panels ist dieses Problem aber wenigstens abmilderbar: Personen können bei genügend hohem Aufwand auch dann für Online-Befragungen gewonnen werden, wenn sie eigentlich nicht das Internet verwenden. Solche Personen müssen dann mittels Brief oder Anruf zur Befragung eingeladen werden. Die Möglichkeit mit relativ geringem Aufwand einen internetfähigen Computer zu erreichen, sind mittlerweile gegeben. Und wenn die Anreize gross genug sind, ist es sogar möglich, computeraverse Personen zu motivieren.

Die Möglichkeiten für Befragte ohne entsprechende Ausstattung einen Computer erreichbar zu machen, sind unterschiedlich. Das niederländische Online-Panel CentERdata stellt diesen Personen sogar einen Computer zur Verfügung (Scherpenzeel, 2008) andere bezahlen besondere Anreize, um die Befragten zu bewegen, selbstständig einen Zugang zum Internet zu schaffen (Beispielsweise über Bekannte, Internetcafes oder den Arbeitsplatz).

Nonresponse Mindestens unter den Bedingungen typischer Befragungen in den Sozialwissenschaften (also bei denen weder Zwang eingesetzt werden kann noch z. B. beliebig grosse Anreize zur Teilnahme gesetzt werden können) scheint die Annahme unrealistisch zu sein, dass alle Befragte eine Wahrscheinlichkeit grösser als Null haben teilzunehmen.

Wenn man allerdings als Gedankenexperiment dem renitentesten Befragungsverweigerer unendlich oft eine Einladung zu einer Befragung stellt, scheint es sicher zu sein, dass der Befragte irgendwann in einem dieser unendlich vielen Versuche bereit ist, zumindest einige Fragen zu beantworten. Er hat also eine sehr kleine Teilnahmewahrscheinlichkeit, aber doch eine grösser als Null.

Es ist wünschenswert, dass die Teilnahmewahrscheinlichkeiten für alle Personen hoch sind. Es gibt sehr viele Einflussvariablen, die bestimmen, wie hoch die Teilnahmequoten sind. Wenn alle in der Literatur vorgeschlagenen Möglichkeiten ausgeschöpft werden, können sehr hohe Teilnahmeraten erzielt werden. Der Anteil derer, die dann noch verweigern, ist dann sehr klein. Für einen Überblick siehe
Groves und Couper (1998), Dillman (2000), Groves et al. (2002) oder Groves et al. (2004).

3.3 Schätzung der Teilnahmeneigung

3.3.1 Parametrische Schätzung

Der Zusammenhang zwischen den *propensity score* ρ_i und X_i wird oft mittels generalisiertem linearem Modell charakterisiert werden (siehe z. B. Alho (1990), Folsom (1991) oder Ekholm und Laaksonen (1991)) und hat dann die Form

$$g(\rho) = X^{\mathsf{T}}\beta, \tag{3.10}$$

wobei $g(.)$ eine Linkfunktion ist. Schouten et al. (2009) nehmen an, dass es sich um eine logistische Regression handelt. Gleichung 3.10 hat dann die Form

$$log\left(\frac{\rho_i}{1-\rho_i}\right) = X_i^{\mathsf{T}}\beta \tag{3.11}$$

Um ein solches *generalized linear model* rechnen zu können, muss es eine genügend grosse Anzahl Antwortausfälle geben, die bezüglich X_i variieren. Die statistischen Eigenschaften einer parametrischen Schätzung von ρ werden von Kim und Kim (2007) diskutiert.

3.3.2 Klassifikations- und Regressions-Bäume

Eine im Vergleich zu parametrischen Methoden sehr viel flexiblere Methode die Zugehörigkeit von Objekten zu Klassen zu prognostizieren, sind Klassifikations-Bäume (*regression-trees*, auch *decision trees* oder ausführlicher *Classification and regression trees (CART)*). Für eine Einführung siehe z. B. Ripley (1996) oder Hastie et al. (2009). Mit Hilfe eines Baumes können bezüglich der unabhängigen Variablen Regionen von Beobachtungen bestimmt werden, die ähnliche Werte bei der abhängigen Variable haben. Bäume sind eine Menge von hierarchischen Regeln, die auf den unabhängigen Variablen basieren. Baummodelle wurden zuerst entwickelt von Morgan und Sonquist (1963), die sie auch in die Sozialwissenschaften eingeführt haben. Heute spielen baumbasierte Modelle in den Sozialwissenschaften nur eine untergeordnete Rolle, sondern werden eher in der Botanik und Medizin angewendet (Ripley, 1996). Breiman et al. (1984) haben auch durch die Entwicklung neuer Methoden Baummodelle revitalisiert.

Sprache der Bäume ist graphisch

Baummodelle werden typischerweise graphisch präsentiert, ein Beispiel ist Abbildung 7.1 (S. 123). Ein Baum besteht aus einem *root* genannten Anfangsknoten und verästelt sich über weitere Knoten bis in die Blätter (*leafs*). Jeder Knoten, der kein Blatt ist, enthält die Information zu einem Split. Jedes Element wird einem der Blätter zugeordnet. Bäume können daher als hierarchische Klassifizierung von Elementen gesehen werden. Ein Baum wird dadurch konstruiert, dass nacheinander an den Knoten Splits vorgenommen werden, bis jedes Element korrekt klassifiziert werden kann. Zu den Details der Baumkonstruktion siehe Ripley (1996) und Breiman et al. (1984).

Ähnlich wie bei einer Regression gilt es, die Summe der quadrierten Resi-

duen zu minimieren. Dafür wird zunächst eine Variable und ein Teilungskriterium gesucht, um so die Observationen in zwei Gruppen einzuteilen. Dabei wird unter allen Variablen diejenige Kombination gewählt, welche die stärkste Reduktion der Residuen-Quadartsumme ergibt. Dann wird für jede der beiden Gruppen (auch Äste genannt) die nächste Variable und ein passendes Teilungskriterium gesucht, die wieder eine möglichst gute Teilung der jeweiligen Gruppe bezüglich der abhängigen Variable ermöglicht wird. Ein solcher Teilungspunkt wird Knoten genannt. Diese Teilung wird fortgeführt, bis ein Abbruchkriterium erreicht ist. Ein typisches Abbruchkriterium ist die Mindestgrösse von fünf Observationen pro Blatt, wobei ein Blatt der «unterste» Ast ist, also einer, der nicht weiter geteilt wird.

Handelt es sich bei den zur Entwicklung des Baums zur Verfügung stehenden Elementen um eine Teilmenge der Elemente, über die eine Aussage getroffen werden soll –es sich also insbesondere um eine Stichprobe handelt– darf die Verästelung des Baums nicht zu fein sein. Die Entwicklung (*growing*) des Baums muss abgebrochen werden, bevor alle Elemente klassifiziert sind, um sogenanntes Rauschen (*noise*) zu vermeiden. Es käme sonst zum *overfitting*, würde der Baum dann auf weitere Elmente der Grundgesamtheit angewendet werden. Diese Logik entspricht der die auch bei der Modellierung von Regressions-Modellen mit Vorwärts- und Rückwärtseliminierung bekannt ist. Diesen Abbruch der Baumentwicklung wird Stutzen (*pruning*) genannt.

Stutzen

Der zunächst entwickelte Baum, der die Zuordnung aller zur Verfügung stehender Elemente erlaubt, wird daher gestutzt (*pruning*), um eine solche Überanpassung zu verhindern. Als Kriterium einen Baum zu stutzen, wird nicht ein absoluter Werte festgelegt (i. S. v. beispielsweise einer Regel der Art «mehr als x Verästelungen sind nicht erlaubt»), sondern dies geschieht mittels eines Glättungsparameters (*smoothing parameter*). Im vorliegenden Fall wurde das so genannte *cost-complexity criterion CC* zur Bestimmung der Baumtiefe benutzt (Breiman et al., 1984). CC ergibt sich aus dem Grad der Missklassifikation plus einer Strafe für die Anzahl von Endknoten.

$$CC = \sum RSS_i + \lambda x \tag{3.12}$$

RSS_i ist die Summe der quadrierten Residuen (*residual sum of squares*) für alle Endknoten i. x ist die Anzahl der Endknoten und λ ein willkürlich

wählbarere Glättungsparameter. Je kleiner λ ist, desto eher werden grosse Bäume bevorzugt und umgekehrt.

Ein Baum ist dann gut angepasst, wenn er die tatsächlichen Zusammenhänge zwischen der abhängigen Variable und den unabhängigen widerspiegelt und nicht nur für den vorliegenden Datensatz eine optimale Anpassung bietet. Um dies zu überprüfen und um einen geeigneten Wert für λ zu finden, wird ein Baum kreuz-validiert (*cross-validated*). Dazu wird der Datensatz in k zufällige Gruppen eingeteilt. Typische Werte für k sind 5 oder 10, wir haben 10 gewählt. Nacheinander wird jede der k Gruppen ausgewählt. Mit den verbleibenden 9 Gruppen wird dann ein Baum entwickelt und dann eine Prognose für die ausgewählte Gruppe vorgenommen. Im Anschluss kann die durchschnittliche Fehlklassifikationsrate für alle k Gruppen berechnet werden. Das Ziel ist es, dass diese durchschnittliche Fehlklassifikationsraten so klein wie möglich sind.

ρ schätzbar als $\hat{p}_{mk}$

Für jedes Blatt m, welches eine Region R_m mit N_m Beobachtungen repräsentiert, gilt im Falle eines Klassifikations-Baums $k(m) = \arg\max_k \hat{p}_{mk}$. Es wird also für alle Beobachtungen eines Blattes der Wert prognostiziert, der innerhalb dieser Gruppe von Beobachtungen am häufigsten vorkommt, da $\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i = k)$ den Anteil aller Beobachtungen angibt, die den Wert k in der abhängigen Variable haben. Quantitative Variablen müssen zunächst kategorisiert werden. Soll wie in diesem Fall hier mittels Baum eine Wahrscheinlichkeit prognostiziert werden, dass eine Beobachtung i zu einer Klasse $y_i = 1$ gehört, dann bestimmt sich die Wahrscheinlichkeit unmittelbar als $\hat{p}_{mk}$.

Gütekriterien

Drei verschiedene (wenngleich verwandte) Kriterien zur Bestimmung der Güte eines Baums sind gebräuchlich:

Cross-Entropie, Devianz	$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$
Fehlklassifikationsrate	$\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk}(m)$
Gini-Index	$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk'})$

Alle drei Kriterien bilden einen ähnlichen Sachverhalt ab, da die Fehlklassifikationsrate aber am anschaulichsten ist, beschränken wir uns auf ihre Darstellung.

3.3.3 Weitere nicht-parametrische Verfahren

Kernel smoother

Giommi (1984) schlägt vor, einen *kernel smoother* (Nadaraya-Watson Schätzer) zu verwenden, um die Antwortwahrscheinlichkeiten zu schätzen. Da Silva und Opsomer (2006, 2008, 2009) zeigen, dass mit dieser Methode der Populationsmittelwert konsistent geschätzt werden kann. Sie erweitern allerdings den Ansatz und schlagen stattdessen vor, eine *local polynomial regression (LOESS)*[4]. zu verwenden. Sie argumentieren (und zeigen an einem exemplarischen Datensatz), dass die lokale Schätzung der *Propensity*-Funktion zu einer besseren Annäherung an die tatsächliche, zugrunde liegende Funktion ermöglicht.

LOESS

Bei einer LOESS Schätzung wird die Anpassung des Modells nicht global für alle Beobachtungen vorgenommen. Diese werden vielmehr in Segmente aufgeteilt, für die dann pro Segment eine Modellanpassung vorgenommen wird. Technisch wird für jeden Datenpunkt ein Polynom angepasst, zu dessen Schätzung ein Subset der Daten in unmittelbarer Nachbarschaft des Datenpunktes verwendet wird. Dazu wird die Methode der gewichteten kleinsten Quadrate (weighted least squares method) verwendet, wobei die Datenpunkte ein umso höheres Gewicht bekommen, je näher sie beim eigentlichen Datenpunkt liegen. Es wird also für jeden Datenpunkt eine eigene Regression gerechnet. Zu den technischen Details von LOESS siehe Cleveland (1979) und Cleveland und Devlin (1988).

Vom Anwender müssen drei Dinge festgelegt werden: der Grad, den das Polynom annehmen darf, die Bandbreite an Daten, die zur Berechnung der lokalen Regressionen herangezogen werden sollen, und die Verteilung der Gewichte innerhalb dieses Bereichs. Diese Entscheidungen haben eine deutliche Auswirkung auf die zu schätzende Funktion (Cleveland und Devlin, 1988).

- Wird z. B. die Bandbreite des zu berücksichtigenden Bereichs an Daten (dem sog. *smoothing-parameter a*) zu klein gewählt, wird von der Funktion auch der Zufallsfehler erfasst. Ist der Bereich zu gross, können lokale Phänomene übersehen werden.

- Der Grad des Polynoms wird typischerweise auf eins oder zwei gesetzt

4 Auch *locally weighted scatterplot smoothing* genannt

(beträgt er Null, wird aus LOESS ein gleitender Mittelwert). Höhergradige Polynome bergen wieder die Gefahr, dass es zu einem *overfitting* kommt, also zu viele Zufallsfehler berücksichtigt werden.

- Obwohl im Prinzip viele verschiedene Gewichtungsfunktionen möglich sind, wird typischerweise die tri-kubische Funktion verwendet:
  $w(x) = \left(1 - |x|^3\right)^3 I\,[|x| < 1]$

LOESS ist auf Grund der vielen je nach Befragung zutreffenden Entscheidungen weniger geeignet, automatisiert angewendet zu werden. Soll das *propensity score adjustment* weitestgehend automatisiert angewendet werden, ist LOESS daher keine sehr praktikable Methode.

3.4 Propensity Scores bei Web-Panel-Befragungen

Die Idee, *propensity scores* auch bei Online-Panels als Mittel zur Gewichtung einzusetzen, wurde zuerst beim Marktforschungsinstitut *Harris Interactive (HI)* entwickelt (Taylor, 2000; Schonlau et al., 2004; Lee, 2006b). Allerdings ist das Vorgehen nicht genau dokumentiert und es ist nicht bekannt, welche Kovariate von Harris Interactive benutzt werden. (Vergleiche S. 33).

Referenzbefragung

Um *propensity scores* berechnen zu können, müssen – wie geschrieben – externe Informationen vorliegen. Insbesondere muss Klarheit über die Verteilung der Kovariate $\mathbf{X}$ in der Population auf der Ebene der Individuen bestehen. Es existieren mindestens bei allgemeinen Bevölkerungsumfragen aber keine Variablen, deren Ausprägung für alle Individuen der Population bekannt ist. Die Verteilung von X muss geschätzt werden. Praktisch heisst dies, dass eine Referenz-Befragung durchzuführen ist, die diese Kovariate erhebt. Diese Befragung muss von einer möglichst hohen Qualität sein, da sie die Grundlage der Gewichtung ist. Selektionsfehler und Bias, die in dieser Referenzbefragung enthalten sind, werden sonst auch für die Online-Befragung übernommen. In der Praxis ist es allerdings sehr schwierig, eine Befragung zu realisieren, bei der es nicht zu *nonignorable nonresponse* kommt.

Es gilt daher, all die in der Literatur beschriebenen Hinweise zur Durch-

führung einer Befragung mit höchster Qualität einzuhalten[5]. Insbesondere ist es notwendig, eine höchstmögliche Teilnahmequoten zu erreichen, damit die Wahrscheinlichkeit steigt, dass auch solche Personen in der Befragung enthalten sind, die eine niedrige Teilnahmewahrscheinlichkeit haben.

Dass diese Mühe nicht notwendigerweise zum Erfolg führt, zeigen z. B. Sidenvall et al. (2002). Sie haben einen sehr hohen Aufwand betrieben, Verweigerer doch noch zur Teilnahme zu motivieren (Anrufe, Schreiben, persönliche Besuche und dergleichen). Es zeigt sich, dass die Gruppe der hinzugewonnenen Befragten sich kaum oder gar nicht von denjenigen unterscheidet, die schon in der ersten Welle teilgenommen haben. Aber unter Umständen können für eine erfolgreiche Modellierung der Teilnahmewahrscheinlichkeit schon die ganz wenigen «Abweichler» entscheidend sein.

Die in der Referenz-Befragung gesammelten Informationen können als Grundlage für die Gewichtung der eigentlichen Online-Befragungen herangezogen werden. Handelt es sich um laufende Panel-Befragungen, ist es möglich, die Referenz-Befragung nur in zeitlich grossen Abständen zu wiederholen. Wie gross diese Abstände sein dürfen, muss empirisch ermittelt werden und hängt davon ab, wie schnell sich die Kovariaten in der Grundgesamtheit ändern. Es empfiehlt sich daher, dass als Kovariate möglichst stabile Merkmale gewählt werden. Wird für den Panel laufend (offline) rekrutiert, ist es auch möglich, die Rekrutierung als Referenzbefragung zu definieren. In diesem Fall müssen vor der entscheidenden Frage, ob die Person zu einer Teilnahme am Panel bereit ist, die Kovariaten erhoben werden.

Praktisches Vorgehen, Kovariate zu erheben

Die Fragen, die die Kovariate bilden und in der Referenz-Befragung erhoben wurden, müssen dann bei jeder Online-Panel-Befragung mit erhoben werden. Der Aufwand, den diese Fragen verursachen, darf also nur sehr gering sein. Ausserdem darf es bei den Kovariaten möglichst nicht zu fehlenden Werten kommen. Sollte dies dennoch geschehen, müssen diese imputiert werden[6].

5 Nochmals sei auf Groves und Couper (1998), Dillman (2000), Groves et al. (2002) oder Groves et al. (2004) als gute Einstiegsliteratur in das Thema verwiesen.

6 Bei Online-Befragungen besteht natürlich noch die technische Möglichkeit, Fragen zu Pflichtfragen zu erheben, d. h. dass Befragte nur dann die eigentlichen Fragen der Erhebung präsentiert bekommen, wenn sie die Fragen zu den Kovariaten vollständig beantwortet haben. Es ist zu diesem Zeitpunkt noch nicht klar, welches Vorgehen in der Praxis besser geeignet ist

Ist einmal klar und festgelegt, welche Kovariate benutzt werden sollen, ist es nicht mehr nötig, für jede Befragung neu zu modellieren. Es muss nicht jedes Mal neu entscheiden werden, ob gegebenenfalls nur ein Teil der der Kovariate verwendet werden soll, da z. B. einige keinen Einfluss in der logistischen Regression haben. Die Bestimmung der Gewichte kann automatisiert erfolgen.

Rubin und Thomas (1996) zeigen, dass es sogar besser sein kann, Kovariate beizubehalten, auch wenn diese nicht signifikant die interessierende Variable Y vorhersagen können. Ausserdem ist es häufig nicht problematisch, wenn das Modell misspezifizert ist, wie Drake (1993) in ihren Simulationen zeigen kann. Der resultierende Bias aus einem schlecht spezifizierten Modell ist demnach sehr viel geringer als der Bias ohne ein Modell benutzt zu haben. Das vereinfacht ein automatisiertes Vorgehen. Hulliger et al. (1997) hat für die Schweizerische Arbeitskräfte Erhebung eine Panel-Attrition Gewichtung entwickelt. Dabei zeigte es sich, dass gewisse Variablen in einem Jahr nicht nötig gewesen wären, aber im darauf folgenden Jahr wieder signifikant wurden. Das ist auch beim PSA zu erwarten: da wie geschrieben ρ_i nicht nur von X_i abhängt, sondern auch von Y_i abhängen kann und diese Y_i bei jeder Befragung unterschiedlich sind, ist es zu erwarten,dass mindestens einige Kovariate bei jeder Befragung eine unterschiedliche Erklärungskraft haben. Dies ist aber unproblematisch. Wenn die Kovariate keinen oder keinen hohen Einfluss im modell haben, sollte der resultierende Schätzer durch sie nicht verzerrt werden. Es kann sich lediglich die Varianz des Schätzers erhöhen. Das ist natürlich auch nicht erstrebenswert, zu Gunsten eines vereinfachten Vorgehens aber akzeptabel. Letztendlich muss in der Praxis für jede Befragung entschieden werden, ob sich der Aufwand lohnt, neu zu modellieren oder ob das automatisierte Vorgehen ausreichend ist.

Es ist möglich, dass es zu aufwändig und teuer sein kann, die Kovariaten bei jeder Befragung mit zu erheben. Trotzdem kann es hilfreich sein, sie mindestens auf der Ebene des Panels erhoben zu haben. Erstens besteht die Möglichkeit, die Teilnahmewahrscheinlichkeit für jede befragte Person zu modellieren, da die Kovariate bereits erhoben wurden. Da die Kovariaten auf individueller Ebene aber nicht notwendigerweise zeitlich stabil sind, ist es auf jeden Fall zu empfehlen, diese bei jeder Befragung neu zu erheben.

R-Indikator

Werden die Kovariaten auf Panelebene erhoben, besteht ausserdem die Möglichkeit, das Panel selbst besser zu überwachen. Insbesondere ist

es einfacher, einen möglichen biasbehafteten Panelschwund zu erkennen. Dies kann mit Hilfe von R-Indikatoren geschehen. R-Indikatoren sind ein Indikator für die Repräsentativität einer Befragung (bzw. eines Panels) und basieren auch auf Kovariaten, die die Teilnahmewahrscheinlichkeit erklären sollen. Diese R-Indikatoren können auch in einem zweiten Schritt helfen, die Repräsentativität der eigentlichen Befragung zu schätzen und dadurch bei der Entscheidung helfen, ob ein *propensity score adjustment* notwendig ist. Verschiedene Möglichkeiten R-Indikatoren zu bestimmen, sollen im Folgenden vorgestellt werden.

3.5 R-Indikatoren auf Basis der Teilnahmewahrscheinlichkeit

Der am häufigsten kommunizierte Indikator für die Güte von Befragungen ist die Teilnahmerate. Es wird offensichtlich die Annahme unterstellt, dass je höher die Teilnahmerate ist, desto besser sei auch die Repräsentativität. Im folgenden Abschnitt 3.5.1 wird gezeigt, dass die Berechnung der Teilnahmeraten bei Web-Panel-Befragungen willkürlich und daher meist nicht informativ ist[7]. Die anschliessenden Abschnitte zeigen bessere, alternative Indikatoren auf der Basis der geschätzten Teilnahmewahrscheinlichkeit.

Verwandschaft zu PSA

Eine Voraussetzung, um solche R-Indikatoren zu konstruieren, ist das Vorhandensein externer Informationen auf der Ebene der Stichprobe. Damit ist gemeint, dass gewisse Informationen sowohl für die Respondenten wie auch die Verweigerer vorliegen müssen, die Verwandschaft zu PSA ist offensichtlich. Im Falle von Web-Panel-Befragungen ist das kein Problem, wenn man als Bezugspopulation das Panel betrachtet und nicht die Gesamtbevölkerung. Dann lassen sich bei der Rekrutierung und bei der eigentlichen späteren Befragung Fragen einbauen, die einen Vergleich ermöglichen können. Ist die Bezugspopulation (korrekterweise) allerdings die angestrebte Grundgesamtheit der Befragung, also typischerweise die Gesamtbevölkerung, dann ist es schwieriger, diese Informationen auf Ebene der Stichprobe zu erhalten.

Es gibt unterschiedliche Vorschläge, wie mit Hilfe der externen Informatio-

7 Dass es auch bei anderen Befragungsmodi, sogar in der amtlichen Statistik und anderen eigentlich qualitativ ambitionierten Befragungen aus der Wissenschaft bei deren Berechnung oft recht exotisch zugeht, illustriert Smith (1995).

nen R-Indikatoren konstruiert werden können, die in den Abschnitten 3.5.2 und 3.5.3 besprochen werden.

3.5.1 Exkurs: Willkürliche Berechnung der Teilnahmeraten

Befragungen über das Internet und speziell über Web-Panels durchzuführen, ist noch ein relativ junges Phänomen. Es gibt daher noch keine einheitliche Art, Teilnahmeraten zu berechnen (Callegaro und Disogra, 2008). Bei «konventionellen», d. h. nicht web-basierten Befragungen, ist die Bestimmung der Teilnahmeraten ein vieldiskutiertes Thema (Groves und Couper, 1998; Schnell, 1997; Schafer und Graham, 2002). Dabei ist grundsätzlich aber ziemlich klar, wie Teilnahmequoten zu berechnen sind: nämlich als Anteil derer, die an einem Interview teilgenommen haben, an allen, die kontaktiert wurden (Groves et al., 2002).

Bei Web-Befragungen ist dies nicht mehr klar[8]. Im Falle von Opt-In-Befragungen ist die Angabe der «Teilnahmequote» trivialerweise nicht möglich, da es den Nenner «alle, die kontaktiert wurden» nicht gibt. Häufig ist bei solchen Befragungen die angestrebte Grundgesamtheit die Menge aller Personen, die im Befragungszeitraum die Seite mit der Bannerwerbung besucht haben. Es existieren aber verschiedene technische Hilfsmittel, um Werbung bei der Darstellung von Webseiten zu unterdrücken[9]. Da nicht klar ist, wie viele Besucher der Seite ein solches Tool verwenden, ist folglich auch nicht bekannt, wie viele Personen die Werbung überhaupt potentiell gesehen haben könnten. Einfach die Klicks auf der Seite zu zählen und dabei die IP-Adresse als Identifikator zu speichern, ist nicht ausreichend. Details zu den hier nicht weiter interessierenden Opt-In-Panels siehe Postoaca (2006).

Teilnahmeraten bei Panels

Auch bei Web-Befragungen, bei denen die Auswahl der Befragten aus einem Panel erfolgt, ist nicht klar, was genau mit «Teilnahmerate» gemeint ist. Dies können sein (wobei die berichtete Teilnahmerate unter gleichen Bedingungen von oben nach unten zunimmt):

1. Anteil aller Befragten, die die Befragung vollständig ausgefüllt haben

8 Natürlich ist es auch bei anderen Interviewmodi oft nicht trivial, die Teilnahmequoten zu bestimmen. Ein Beispiel, wie Teilnahmequoten bei RDD-Interviews informativer bestimmt werden können, als der Anteil derjenigen, die sich an der Befragung beteiligt haben, an allen, die kontaktiert wurden, findet man in Ezzati-Ricea et al. (2000).

9 Das populärste Tool ist das Plugin *AdBlock Plus* für die Browser Firefox und Safari.

und keine *frauds* sind (also –vermutlich– richtige Antworten gegeben haben), an allen Kontaktierten.

2. Anteil derer, die die Befragung vollständig ausgefüllt haben, an allen Kontaktierten.

3. Anteil derer, die höchstens eine akzeptable Menge an Fragen nicht vollständig ausgefüllt haben, an allen Kontaktierten.

4. Anteil der Befragten, die die letzte Seite erreicht haben, an allen Eingeladenen.
   Das entspricht dem Anteil aus 3. plus allen, die sich nur durch die Befragung geklickt haben und höchstens die persönlichen Angaben ausgefüllt haben, um die *Incentives* zu bekommen. Solche Personen werden auch *lurker* genannt.

5. Anteil der Befragten, die mindestens die ersten Fragen ausgefüllt haben, an allen Kontaktierten.

6. Anteil der Befragten, die mindestens die erste Seite der Befragung aufgerufen haben, an allen Kontaktierten.

Diese Wege, die Teilnahmequote zu messen, betreffen nur den quantitativen Vergleich zwischen denen, die kontaktiert wurden, zu allen, die geantwortet haben. Bei Panel-Befragungen ist diese Angabe aber unvollständig.

Hinzu kommt, dass es auch schon bei der Auswahl zum Panel Nonresponse und Verweigerer zur Teilnahme am Panel gab. Beide Ausfallprozesse schliessen schon vor dem Beginn der eigentlichen Web-Befragung sehr viele angestrebte Befragte aus und bewirken einen Bias. Die Teilnahmebereitschaft bei Telefoninterviews wird bei LINK mit rund 60 % angegeben. Von diesen erklären sich rund 10 %–20 % bereit, am Panel teilzunehmen. Ähnliche Zahlen berichten auch andere Panelbetreiber, wie CentERdata aus Holland (Scherpenzeel, 2008).

Standards

Bisher hat sich noch kein einheitlicher Standard zur Berechnung der Teilnahmeraten durchsetzen können. So empfehlen verschiedene wichtige Journale, wie z. B. dem (neuerdings sehr auf Web-Surveys orientierten) *Public Opinion Quarterly* und dem *International Journal of Public Opinion Research* die Verwendung des traditionellen AAPOR-Standards zur Berechnung der

Teilnahmequoten bei Web-Befragungen. Allerdings beschränkt sich diese Empfehlung auf Befragungen von Personen, die in einer Liste enthalten sind (AAPOR, 2008)[10]. Bei Panelbefragungen ergibt sich die Teilnahmequote daher als der Anteil derjenigen, die an einer Befragung teilgenommen haben, an allen, die eingeladen wurden, gemäss einer der oben vorgeschlagenen Definitionen. Andere Journale, wie z. B. das *Journal of Medical Internet Research*, empfehlen die Vermeidung von Begriffen wie «Response Rate» und empfehlen die Angabe von *completition rates* und *view rates* (Callegaro und Disogra, 2008), aber auch, ohne diese irgendwie zu präzisieren.

Verschiedentlich wurde als Erweiterung traditioneller Teilnahmeraten die Berechnung einer kumulativen Teilnahmequote vorgeschlagen (Huggins und Eyerman, 2001; Tourangeau, 2003; Couper, 2007; Callegaro und Disogra, 2008). Eine vollständige Teilnahmequote REP^{VOLL} ergibt sich demnach multiplikativ aus der Teilnahmequote bei der Rekrutierung REP^{REK}, dem Anteil derer am Rekrutierungsinterview, die sich bereit erklären am Panel teilzunehmen REP^{PAN1}, dem Anteil derjenigen, die zum Zeitpunkt der Befragung noch aktive Panelisten sind REP^{PAN2} und dem Anteil derjenigen an allen Eingeladenen, die vollständig und richtig geantwortet haben REP^{REAL}. Siehe dazu auch nochmals Abbildung 2.1, auf S. 22. In der Grafik nicht erwähnt sind weitere mögliche Ausfälle. Z. B. werden häufig nicht alle Personen zum Online-Panel zugelassen, beispielsweise werden gelegentlich Alterslimite vorgeschrieben oder eine Mindestnutzung des Internets vorausgesetzt.

Panelabrieb

Wie erwähnt ist es schwierig –oder sogar unmöglich– die Panelmortalität zu quantifizieren. REP^{PAN2} lässt sich daher nie eindeutig bestimmen, sondern allenfalls nur langfristig schätzen. Selbst wenn man Teilnahmeraten einheitlich berechnet, ist nicht klar, inwieweit die Angabe tatsächlich bezüglich der Qualität der erhaltenen Daten aussagekräftig ist. Häufig ist die Teilnahmerate bei Panel-Befragungen nur ein Indikator für die Art des Panelmanagements. Williams et al. (2006) reproduzieren eine Offline-Studie, die sie als Benchmark benutzen, mit 19 unterschiedlichen holländischen Web-Panels. Die Teilnahmeraten schwanken zwischen 18 % und 77 %, wobei

10 ESOMAR gibt keine konkrete Empfehlung, wie Teilnahmequoten zu berechnen sind, sonder beschränkt sich auf die Empfehlung «publishes a clear statement of the sample universe defnition used in a given survey, the research approach adopted, the response rate achieved and the method of calculating this where possible» (ESOMAR, 2005), allerdings wird das Dokument momentan überarbeitet, wobei eine präzisere Definition angekündigt wurde.

die Teilnahmeraten in keiner Weise die Qualität der Daten vorhersagen können, gemessen als Vergleich mit den Offline-Daten. Sie kommen zu dem Ergebniss, dass «Response percentage does not indicate sample or panel quality. It reflects a panel business strategy.»

Die Unterschiede in den Teilnahmeraten resultieren allein aus der Entscheidung des Panelmanagements, welche Personen im Panel behalten werden und welche nicht. Die Panels mit den niedrigen Teilnahmeraten führen alle einmal eingeschriebenen Personen als Panelisten (vermutlich auch, um dokumentieren zu können, wie gross ihr Panel ist). Die Panels mit mittleren Teilnahmeraten betreiben eine mehr oder weniger gute Panelpflege und berücksichtigen jeweils die Panelmortalität in ihren Angaben. Die Panels mit den sehr hohen Teilnahmeraten behalten nur solche Personen als Panelisten, die sich durch eine sehr hohe Teilnahmequote in vergangenen Befragungen bewährt haben. Da so nur Hochmotivierte im Panel verbleiben, kann eine hohe Teilnahmequote gelegentlich ironischerweise auf einen starken Bias hindeuten, vorausgesetzt die Motivation korreliert mit der interessierenden Variable.

3.5.2 R-Indikator der RISQ-Gruppe

Wenn die Teilnahmewahrscheinlichkeiten für alle Elemente der Population und der Stichprobe bekannt sind beziehungsweise ‹realistischer› geschätzt werden können, können diese Informationen genutzt werden, um die Repräsentativität einer Befragung zu beurteilen.

R-Indikator ist Varianz der PS

Repräsentativität kann gemessen werden als Varianz der *propensity scores* auf der Ebene der Population (Schouten et al., 2009)[11]. Je grösser diese ist, desto grösser wird auch der Bias sein. Die Standard Abweichung beträgt

$$S(\rho) = \sqrt{\frac{1}{N-1}\sum_{U}(\rho_i - \bar{\rho}_U)^2}, \tag{3.13}$$

wobei $\bar{\rho}_U = \sum \rho_i / N$ das arithmetische Mittel des *propensity scores* ist. Das

11 Die genannten Autoren sowie weitere Mitarbeitende der nationalen statistischen Institute der Niederlande, Norwegen und Slowenien sowie einiger Universitäten arbeiten zusammen im Projekt RISQ (`www.risq-project.eu`), um R-Indikatoren zu entwickeln. Bei RISQ handelt es sich um ein EU-finanziertes Projekt (FP7). Das offizielle Projektende war am 30.06.2010.

theoretische Maximum von $S(\rho)$ beträgt 0.5, da $S(\rho) \leq \sqrt{\bar{\rho}_U(1-\bar{\rho}_U)} \leq 0.5$. Um einen Repräsentativitätsindikator zu konstruieren, der einen Wertebereich von [0,1] hat, wobei 1 keinerlei Repräsentativität und 1 höchste Repräsentativität indizieren soll, definieren Schouten et al. (2009) als Repräsentativitätsindikator

$$R(\rho) = 1 - 2S(\rho). \tag{3.14}$$

Das theoretische Maximum von $R(\rho)$ beträgt immer 1. Das Minimum hängt dagegen vom durchschnittlichen *propensity score*, also der Teilnahmerate ab. Nur wenn $\bar{\rho}_U = 0.5$ gilt, kann das Minimum von 0 erreicht werden. Für $\bar{\rho}_U = 1$ und $\bar{\rho}_U = 0$ ist das Minimum von $R(\rho) = 1$. Das jeweilige Minimum ist definiert als $1 - 2\sqrt{\rho_U(1-\rho_U)}$. Abbildung 3.1 verdeutlicht diesen Zusammenhang nochmals.

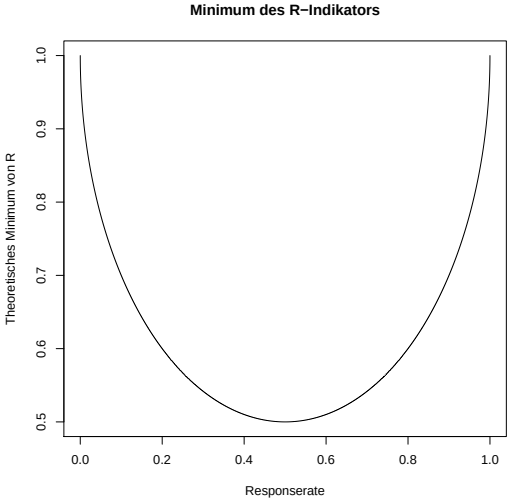


Abbildung 3.1: Minimum des R-Indikators in Abhängigkeit der Teilnahmerate

$R(\rho)$ kann unter Berücksichtigung der Designgewichte d_i geschätzt werden mit

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{S} d_i \left(\rho_i - \hat{\bar{\rho}}_U\right)^2}, \tag{3.15}$$

wobei $\hat{\bar{\rho}}_U$ ein Schätzer für $\bar{\rho}_U$ ist mit

$$\hat{\bar{\rho}}_U = \left(\sum_{S} d_i \hat{\rho}_i\right) / N \tag{3.16}$$

$N - 1$ kann ersetzt werden durch $\sum_S d_i$. Liegen als externe Informationen nur aggregierte Daten vor, kann $R(\rho)$ geschätzt werden als

$$\hat{R}_r(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{r} d_i \hat{\rho}_i^{-1}(\hat{\rho}_i - \hat{\bar{\rho}}_r)^2}, \tag{3.17}$$

wobei gilt

$$\hat{\bar{\rho}}_r = \frac{\sum_r d_i}{N} \tag{3.18}$$

Die RISQ-Gruppe versteht den R-Indikator zusammengefasst als Mass, inwieweit ein –oder genauer gesagt kein– Zusammenhang zwischen den externen Variablen und der geschätzten Teilnahmewahrscheinlichkeit besteht. Der R-Indikator leitet sich technisch aus Cramers V ab. Je geringer die Assoziation zwischen der Teilnahmewahrscheinlichkeit und den Hilfsvariablen ist, desto besser ist die Repräsentativität der Untersuchung.

3.5.3 Särndal und Lundström

Ein ähnlicher Vorschlag für einen R-Indikator stammt von Särndal und Lundström (2008). Auch bei ihnen ist der R-Indikator ein Indikator für nonreponseerzeugten Bias. Allerdings benutzen sie die reziproke Linkfunktion anstatt des Logit-Links im GLM, um die Teilnahmewahrscheinlichkeit zu schätzen.

$$\rho^{-1} = X\beta \tag{3.19}$$

ρ^{-1} nennen sie *influence* und schreiben ϕ. Sie zeigen, dass für eine endliche Population ϕ_i approximiert werden kann mit

$$\phi_{Ui} = x_i^\top \left(\sum_U \rho_i x_i x_i^\top \right)^{-1} \sum_u x_i \tag{3.20}$$

Sie leiten als R-Indikator ab:

$$\mathcal{Q}^2(\rho) = \frac{\sum_u \rho_i (\phi_{Ui} - \bar{\phi}_{\rho U})^2}{\sum_u \rho_i}, \tag{3.21}$$

wobei $\bar{\phi}_{\rho U}$ das ρ_i gewichtete Mittel von ϕ_{Ui} ist, gegeben als

$$\bar{\phi}_{\rho U} = \frac{\sum_u \rho_i \phi_{Ui}}{\sum_U \rho_i} \tag{3.22}$$

Särndal und Lundström schlagen vor, $\mathcal{Q}^2(\rho)$ zu schätzen als

$$q^2 = \frac{\sum_r d_i \, (\hat{\phi}_i - \bar{\phi}_r)^2}{\sum_r d_i}, \tag{3.23}$$

wobei $\hat{\phi}_i$ definiert ist als

$$\hat{\phi}_i = x_i^\top \left(\sum_r d_i x_i x_i^\top \right)^{-1} \sum_s d_i x_i, \tag{3.24}$$

und $\hat{\phi}_r$ definiert ist als

$$\hat{\phi}_r = \frac{\sum_r d_i \hat{\phi}_i}{\sum_r d_i} = \frac{\sum_s d_i}{\sum_r d_i} \tag{3.25}$$

Sind nur aggregierte Daten vorhanden, muss in Gleichung 3.5.3 $\hat{\phi}$ ersetzt werden durch $\bar{\phi}$, definiert als

$$\tilde{\phi}_i = x_i^\top \left(\sum_r d_i x_i x_i^\top \right)^{-1} \sum_U x_i \tag{3.26}$$

3.5.4 Bemerkungen und Empfehlungen

Externe Informationen nötig

Es ist nur dann möglich R-Indikatoren zu berechnen, wenn externe Informationen vorliegen. Im Falle von Web-Panels ist dies kein Problem. Es müssen daher Variablen gefunden werden, mit deren Hilfe die Repräsentativität langfristig gemessen werden kann. Diese Variablen müssen konstant sein in dem Sinne, dass Sie bei der Rekrutierung erhoben werden oder aus anderen Quellen auf Stichprobenebene bekannt sind und immer wieder verwendet werden. Wird das Modell geändert, ergeben sich möglicherweise unterschiedliche Werte für die R-Indikatoren, ohne dass tatsächlich ein Unterschied vorhanden ist.

Empfehlungen

Folgende Empfehlungen können (aufbauend auf Schouten et al. (2009)) zum Gebrauch von R-Indikatoren gegeben werden:

- **Darstellung der Indikatoren**
  - Die Präsentation der R-Indikatoren sollte nicht ohne Angabe der zugrunde liegenden externen Variablen erfolgen.
  - Unterschiedliche Befragungen können nur bezüglich ihrer Repräsentativität verglichen werden, wenn die selben externen Variablen zur Berechnung des R-Indikators verwendet werden.
  - Es ist notwendig, das Konfidenzintervall für den R-Indikator anzugeben.
- **Modellierung der Indikatoren**
  - Es ist zu empfehlen, sowohl das Set der externen Variablen, wie auch mögliche Interaktionstermen zwischen ihnen generell zu fixieren. Werden Variablen verwendet, die nicht mit dem Teilnahmeverhalten korrelieren, hat dies nur einen Einfluss auf den Standardfehler des R-Indikators, nicht jedoch auf seine Höhe.

- Um die *propensity scores* zu berechnen, hat die Wahl der Linkfunktion nur einen kleinen Einfluss auf die R-Indikatoren.

- Die Konfidenzintervalle für die R-Indikatoren können auch bei grossen Stichproben noch sehr gross sein. Bei kleinen Stichproben ist es nicht mehr sinnvoll möglich, R-Indikatoren zu berechnen. Die Grösse der Konfidenzintervalle wird auch nur schwach durch die Anzahl verwendeter externer Variablen beeinflusst.

3.5.5 Kritik an R-Indikatoren

Continuum of resistance

Eine wichtige Annahme, die den R-Indikatoren implizit zugrunde liegt, ist die *continuum of resistance* Hypothese (Lin und Schaeffer, 1995). Sie besagt, dass die Teilnahmewahrscheinlichkeiten auf einem Kontinuum verteilt sind, also stetig sind. Gilt diese Hypothese, dann sollte es so sein, dass die Verweigerer denjenigen in der Stichprobe ähnlich sind, die schwer zu erreichen gewesen sind.

Eine alternative Annahme ist, dass mindestens die Verweigerer in Klassen aufgeteilt werden können, (*classes of nonparticipants model*). Es sind mehrere Klassifizierungen möglich, die verbreitetste ist die Einteilung in zeitweilige Verweigerer (*temporary refusals*), hartnäckige Verweigerer (*hard-to-contacts*) und leicht zu erreichende Personen (*easy respondents*) (Stoop, 2005). Unabhängig davon, wie man diese Klassifizierung genau vornimmt, gilt, dass hier die Verteilung der Teilnahmewahrscheinlichkeiten diskret ist. Wenn die Annahme zutrifft, dass die Wahrscheinlichkeiten stetig verteilt sind, muss nicht mehr gelten, dass diejenigen, die schwer zu erreichen gewesen sind, denen ähneln, die nicht an der Befragung teilgenommen haben.

Dreierklassifizierung

Stoop (2005) führt eine Metastudie zu Nonresponse-Analysen durch und kommt zu dem Ergebniss, dass die oben genannte Dreierklassifizierung am besten bestätigt wird. Sie zeigt weiter, dass sich die Gruppen der leicht zu Motivierenden und die der vorübergehenden Verweigerer sehr ähnlich sind. Eine praktische Schlussfolgerung ist, dass wenn Befragungsleiter (typischerweise unter sehr hohen Kosten) die Teilnehmerquote um einige Prozentpunkte erhöhen, dies kaum Auswirkungen hat, da meist nur vorübergehende Verweigerer rekrutiert werden können. Die Hartnäckigen werden

(fast) nicht erreicht[12].

Das folgende Zahlenbeispiel soll verdeutlichen, warum eine nicht-stetige Verteilung ein Problem bei der Bestimmung der R-Indikatoren sein kann, insbesondere wenn es eine Gruppe hartnäckiger Verweigerer gibt. Eine Grundgesamtheit soll aus 20 Personen bestehen, wobei 15 Personen eine konstante Teilnahmewahrscheinlichkeit von 0.7 haben, die restlichen Personen jeweils eine von 0.001. Auch die interessierende Variable Y ist in beiden Gruppen jeweils konstant. Für die erste Gruppe gilt $Y_i = 100$, für die zweite Gruppe $Y_i = 50$.

Der R-Indikator hat dann einen Wert von $R_{RISQ} = 0.38$ und der Erwartungswert für das artihmetische Mittel $\bar{Y}$ unter der Bedingung, dass man alle 20 Personen zu einer Befragung einlädt und diese mit ihrer entsprechenden Teilnahmewahrscheinlichkeit partizipieren $\bar{Y}_{0.7} = (15 * 0.7 * 100 + 5 * 0.001 * 50)/(0.7 * 15 + 0.001 * 5) = 99.98$. Verändert sich jetzt die Teilnahmewahrscheinlichkeit der ersten Gruppe von 0.7 auf je 0.5, steigt $R_{RISQ} = 0.57$. Der Erwartungswert von $\bar{Y}$ bleibt aber weiterhin praktisch gleich stark verzerrt $\bar{Y}_{0.5} = 99.97$. Der R-Indikator hat also sehr stark reagiert, obwohl der interessierende Schätzer weiterhin (praktisch unverändert) von einem starken Bias beeinflusst wird.

12 Nicht nur Stoop (2005), sondern z. B. auch Groves und Couper (1998) folgern aus diesem Ergebnis, dass es wohl besser ist, Ressourcen anstatt für aufwändige Konvertierungsversuche lieber für andere Formen der Biasreduktion auszugeben.

4 Die richtigen Kovariate

► Eine Bedingung *propensity score adjustment* anzuwenden, ist es, möglichst gut die Teilnahmeneigung schätzen zu können. Neben der richtigen Modellierung ist es insbesondere die Auswahl der Kovariate, also der erklärenden Variablen, die bestimmt, wie gut die Teilnahmeneigung geschätzt werden kann (siehe Abschnitt 3.2.1. In diesem Kapitel sollen verschiedene theoretische Konzepte vorgestellt werden, die die Ableitung geeigneter Kovariate erlauben könnten. Dies soll die experimentelle Befragung in dem Sinne vorbereiten, als dass in diesem Kapitel der Fragebogen entwickelt werden wird. Zunächst werden formale Kriterien genannt, die an solche Kovariate zu stellen sind. Es schliesst sich die Diskussion verschiedener Theorien an, wobei auch gleich empirisch prüfbare Indikatoren für das Teilnahmeverhalten abgeleitet und als Fragen formuliert werden. Im letzten Abschnitt wird ergänzend ein Abriss über Einflussfaktoren auf das Teilnahmeverhalten gegeben, die in der experimentellen Befragung nicht berücksichtigt werden konnten. ◄

Geschichte PSA bei Web-Befragungen

Bisher wurde das *propensity score matching* nur sehr selten verwendet, um eine Biasreduktion bei Online-Befragungen zu erreichen. Zwei Beispiele konnten bisher gefunden werden[1]: Oberski (2006) bestimmt *propensity scores* im Rahmen der Dutch Mobility Study auf der Basis von Fragen, die sich aus dem Thema der Erhebung ergeben haben, wobei es sich insbesondere um demographische Fragen handelt.

1 Die Anwendung der Methode scheint sich insbesondere im Rahmen der Marktforschung immer grösser Beliebtheit zu erfreuen. Auch andere Panelbetreiber verwenden mittlerweile die Methode. Allerdings ist die Dokumentation, wenn sie überhaupt vorhanden ist, jeweils nicht annähernd ausreichend oder informativ. Ein Beispiel aus der Schweiz ist das Web-Panel von *DemoSCOPE*, `www.demoscope.ch`. Zwar gibt DemoSCOPE an, PSA zu verwenden, dokumentiert aber nicht, wie sie vorgehen und welche Variablen sie verwenden.

Schonlau et al. (2002, 2004, 2007) und Lee (2006b) beschreiben das Vorgehen des NewYorker Markt- und Meinungsforschungsinstituts *Harris Interactive, (HI)*[2]. HI hat nach eigenen Angaben eine Reihe von Fragen identifiziert, die die Bestimmung der *propensity scores* erfolgreich ermöglichen[3]. Im HI Jargon werden solche Fragen *Webographic Questions* genannt. Es handelt sich explizit nicht um traditionelle demografische Fragen, da die enthaltenen Informationen über das Anwortverhalten bei Online-Befragungen nicht als ausreichend betrachtet werden. Welche Fragen verwendet werden, dokumentiert HI nicht (sie werden als Betriebsgeheimnis behandelt). Trotzdem lassen sich die Webographic Questions in HI Befragungen als solche identifizieren, die offensichtlich nicht zum Thema der Befragung passen. Schonlau et al. (2007) nennen als ein Beispiel für solche Fragen «Do you know someone gay?» und Bethlehem (2010) «Do you feel sometimes lonely?».

Warum nochmals neu?

Das Vorgehen von HI ist –jedenfalls soweit es publiziert ist– kritisch zu hinterfragen. Zunächst handelt es sich beim Panel von HI um ein Opt-In Panel von Freiwilligen[4]. Es ist unklar, welche und wie viele Fragen zum PSA verwendet werden. Es existieren kaum Informationen zur Referenzbefragung, mit deren Hilfe die *propensity scores* geschätzt werden. Es ist nur bekannt, dass es sich um eine sehr kleine RDD Befragung handelt (Bethlehem, 2007). Zudem sind unabhängigen Versuche, mit Hilfe der «Webographic Questions» (sofern sie als solche zu erraten sind) eine Biasreduktion zu ermöglichen, mindestens in Holland gescheitert (Bethlehem, 2007).

Unabhängig davon, dass die Fragen von HI nicht übernommen werden können und sollen, ist nicht klar, ob diese Fragen in der Schweiz brauchbare Indikatoren für das Antwortverhalten wären. Das Antwortverhalten bei Befragungen ist immer auch abhängig vom kulturellen Kontext. So zeigen Befunde der kulturvergleichenden Psychologie, dass es kulturelle Kontexteffekte im Verhalten der Befragten gibt. Für einen Überblick siehe Fiske et al.

2 `www.harrisinteractive.com`

3 Allerdings scheint es so zu sein, dass das propensity Score Matching nicht in allen Fällen, jedenfalls nicht in allen Ländern erfolgreich eingesetzt wird. Mindestens in Deutschland hat es Versuche von HI gegeben, *propensity score adjustment* einzusetzen, was aber in nicht geglückt ist (Drewes, 2010).

4 Jedenfalls gilt dies für das Panel der amerikanischen Muttergesellschaft. Der deutsche Ableger wirbt unklar mit einem «Rekrutierungsmix» (`www.harrisinteractive.de/onlinepanel.asp`, Stand 1. Nov. 2010). Was das bedeutet und wie rekrutiert wird, dokumentiert HI Deutschland nicht. Sicher ist nur, dass sich die Panelführung von der der Muttergesellschaft unterscheidet. Insbesondere kommt in Deutschland kein PSA zur Anwendung (Drewes, 2010).

(1998).

Da es noch keine Erfahrungen mit der Auswahl möglichst geeigneter Fragen gibt, müssen diese nach Plausibilitätsgründen ausgewählt und dann empirisch getestet werden. Die Auswahl der Kovariate wäre unproblematischer, wenn es eine einheitliche Theorie zur Erklärung des Teilnahmeverhaltens geben würde. Das ist aber nicht der Fall (Groves et al., 2002; Schnell, 1997). Es muss bei der theoretischen Herleitung Rückgriff auf allgemeine Handlungstheorien genommen werden. Verschiedenste Handlungstheorien können herangezogen werden, um das Teilnahmeverhalten zu klären. Prinzipiell eignen sich alle individualistischen Theorien (Opp, 1976), die empirisch prüfbar sind.

Das Ziel der Arbeit ist es nicht, sich für eine der alternativ vorgeschlagenen allgemeine Handlungstheorien auf Grund theoretischer Überlegungen (d. h. Präferenzen) zu entscheiden.[5]. Der hier gewählte Ansatz soll vielmehr ein pragmatischer sein.

In einem ersten Schritt gilt es, möglichst geeignete Kovariate vorzuschlagen. Dies soll im nächsten Abschnitt 4.1 geschehen. Aus den vorgeschlagenen Variablen müssen dann diejenigen ausgewählt werden, mit deren Hilfe die Teilnahmewahrscheinlichkeit am besten prognostizierbar ist. Das wird mit Hilfe einer experimentellen Befragung geschehen, die in Abschnitt 5 ab S. 87 beschrieben wird. Das Untersuchungsdesign hat in diesem Zusammenhang Vor- und Nachteile. Der Nachteil ist, dass nur der Ausfallprozess zwischen CATI-Rekrutierung und Panelteilnahme modelliert werden kann, also nicht der ganze Ausfallprozess zwischen Grundgesamtheit und Befragung beschrieben wird. Das macht die Ergebnisse für die Einschätzung des gesamten Bias, dem eine Web-Befragung unterliegt, nur eingeschränkt brauchbar. Allerdings werden mittels CATI auch Personen erreicht, bei denen die Wahrscheinlichkeit der Teilnahme an einer Web-Befragung nur sehr gering ist.

Diese Verkürzung ist allerdings auch ein Vorteil. Über den gesamten Aus-

5 Eine konsistente und allgemein akzeptierten Handlungstheorie fehlt in der Soziologie und ist vermutlich eine ihrer grössten Schwächen. So schreiben z. B. Fehr und Gintis (2007) «sociological theory has not developed a coherent, broadly accepted framework that facilitates cumulative scientific progress and explains the emergent aggregate patterns of social behavior in terms of individuals' preferences, their beliefs, and the social and economic constraints they face. Nor has sociological research developed a parsimonious, empirically grounded view of the basic motivational driving forces of human behavior.»

fallprozess ist in der Literatur bereits mehr bekannt, als hier berücksichtigt werden muss. Zum Beispiel ist es wichtig, bei Nonrespondenten zwischen Nichtkontaktierbaren und bewussten Verweigerern zu unterscheiden (Stoop, 2005). Beide Aufallmechanismen müssen unterschiedlich modelliert werden, da sie tatsächlich ursächlich grundverschieden sind. Bei denen, die bei der experimentellen Befragung zugesagt haben, an der Web-Befragung teilzunehmen, gibt es aber keine Nichtkontaktierbare: alle die zugesagt haben, haben auch gültige E-Mail Adressen angegeben.

Technisch Nichterreichbare

Ob diese E-Mails die Personen tatsächlich erreicht haben, ist zwar nicht abschliessend sicher klärbar (es ist keine E-Mail als unzustellbar retourniert worden), aber diese Möglichkeit sollte nicht sehr ins Gewicht fallen. Es macht keinen Unterschied, ob die E-Mail auch gelesen wurde oder ungelesen -aber absichtlich- gelöscht wurde. Nach dem Rekrutierungsinterview sollte klar sein, dass es sich bei der E-Mail um eine Einladung zu einer Befragung handelt. Erklärungstechnisch sollte es keine Rolle spielen, wann die Entscheidung zur Partizipation gefallen ist, ob vor oder nach dem Lesen. Es ist möglich, dass die Einladungs-E-Mail beim Empfänger als Spam klassifiziert und gelöscht wird. Das ist insbesondere bei sehr sensiblen Spamfiltern der Fall. Technisch ist es so, dass viele Einladungs-E-Mails zu Web-Befragungen vom Befragungstool aus versendet werden. Es ist dabei möglich, als Absenderadresse eine beliebige E-Mail Adresse anzugeben, egal ob dem Absender diese gehört oder nicht. Der Postausgangsserver für die E-Mail stimmt dann nicht mehr mit der Domaine der E-Mail (z. B. `@befragungsinstitut.ch`) überein. Es handelt dann sich um so genanntes *mail-spoofing*. Spam-Filter sind in der Lage *spoof-mails* zu erkennen. Diese werden für gewöhnlich als Spam klassifiziert, um z. B. *phising* zu verhindern, bei dem eine falsche Identität vom E-Mail Absender vorgetäuscht wird, um an sensible Daten zu gelangen.

Allerdings gibt es Server, die auf so genannten *white-lists* stehen und gewisse Postausgangsserver als unbedenklich klassifizieren. Viele Anbieter von Befragungs-Tools stehen auf einer solchen Liste. Allerdings ist es abhängig von der Konfiguration des Posteingangsservers, ob und welche *white-lists* akzeptiert werden. Gerade Unternehmen sind erfahrungsgemäss relativ streng bei der Spamfilter-Konfiguration. Personen, die aus aus technischen Gründen die E-Mail nicht erhalten haben, sind also nicht von denen unterscheidbar, die sich bewusst gegen die Teilnahme entschieden

haben. Beide werden als Verweigerer behandelt und zusammen modelliert.

4.1 Auswahl der Kovariate

Kovariate werden benötigt, um die Responsewahrscheinlichkeiten modellieren zu können. Dabei müssen zwei Aspekte berücksichtigt werden, die zu einem Bias führen können: a) manche Personengruppen werden nicht so leicht mittels Internet erreicht wie andere und b) unabhängig vom Befragungsmedium unterscheidet sich die Wahrscheinlichkeit zu antworten zwischen verschiedenen Personen.

4.1.1 Formale Kriterien

Es muss berücksichtigt werden, dass die ausgewählten Fragen später im Rahmen von Befragungen gestellt werden, die einem anderen Thema gewidmet sind. Ein wichtiges Kriterium, dass die Entscheidung zur Partizipation beeinflusst ist das Thema der Befragung (siehe neben vielen anderen z. B. Kühne und Böhme (2006)). Die Modellierung der Teilnahmewahrschienlichkeit muss also bei jeder Befragung neu vorgenommen werden, was praktisch bedeutet, dass die Kovariaten bei jeder Befragung erneut abgefragt erhoben werden müssen. Da Befragungen aus methodischen Gründen (zunehmender Item-Nonresponse) und meist auch aus finanziellen Gründen nicht zu lang werden dürfen, muss der Aufwand für die Beantwortung der dann eigentlich inhaltlich uninteressanten Kovariaten möglichst klein sein. Die Fragen müssen daher möglichst kurz sein, sie dürfen nicht zu viel Aufwand verursachen und es darf sich nicht um zu viele Fragen handeln.

Wenige Fragen

Je weniger Fragen ausgewählt werden, umso besser sind diese dann in spätere Befragungen einbindbar. Insbesondere bei Web-Befragungen zeigt sich, dass mit zunehmender Länge der Befragung die Qualität der Antworten abnimmt. Zum einen nimmt die Zahl der Nonresponses durch Abbrecher kontinuierlich zu (Bosnjak, 2002; Galesic, 2002). Zum anderen nimmt die Qualität der Antworten ab (Ganassali, 2008; Galesic, 2005). Als illustrierendes Beispiel sei angeführt, dass sich die Zahl der «weiss nicht» Antworten mit zunehmender Fragebogenlänge erhöht (Thomas et al., 2003). Ausserdem nimmt die Anzahl derer, die bei Fragen mit gleichen oder ähnlichen Inhalt konsistent antworten, mit der Fragebogenlänge zu

straight-line-effect
(Herzog und Bachman, 1981).

Kurze Fragen

Auch die Fragen selbst dürfen nicht zu lang sein. Lange Fragen haben ceteris paribus niedrigere Response-Raten als kürzere Fragen (Roszkowski und Bean, 1990). Ein häufig angewendetes Mittel, die Länge von einzelnen Fragen zu reduzieren und unnötige Wiederholungen zu vermeiden, sind Matrix-Fragen. Obwohl Matrix-Fragen den Nachteil haben, dass sich die Zahl der Drop-Outs erhöht (Knapp, 2001), sind sie am besten geeignet, da sie den kognitiven Aufwand durch die einheitliche Skala reduzieren.

Fragen, die Multimediaelemente beinhalten, können von vorne herein ausgeschlossen werden, da sie zur Beantwortung zu viel Zeit benötigen. Bei Textfragen ist es vielleicht einfacher, wenn der Fragenkomplex, der nur der Gewichtung dienen soll, aus einer Matrixfrage mit einheitlicher Antwortskala besteht. Das würde den kognitiven Aufwand für die Befragten und damit die Zeit, die sie zum beantworten benötigen, reduzieren.

Einfache Fragen

Selbstverständlich darf die Beantwortung der Fragen nicht mit einem zu grossen Aufwand verbunden sein. Es verbieten sich Fragen, bei denen das Wissen durch die Befragten nicht sofort abrufbar ist, oder bei denen es Unklarheiten bei der Beantwortung gibt. Fragen, die eine präzise Definition der verwendeten Begriffe oder Konzepte erfordern, sind ungeeignet, genauso wie Fragen mit ungewohntem oder anspruchsvollen Antwortschema[6].

4.1.2 Inhaltliche Kriterien

Obwohl dies eigentlich nötig wäre, soll (wie oben geschrieben) im Folgenden nicht der Versuch unternommen werden, das Teilnahmeverhalten mittels einer allgemeinen Theorie zu erklären. Es können und sollen nur Faktoren gefunden werden, die das Teilnahmeverhalten erklären können, ohne dass sie alle gesamthaft Teil einer konsistenten Theorie sind. Orientierung bei der Auswahl der Kovariate bietet dabei der Rückgriff auf eine umfangreiche Nonresponseliteratur. Für eine Übersicht zur Nonresponseliteratur siehe z. B. Groves et al. (2002); Groves und Couper (1998); Schnell (1997); Vehovar et al.

6 Eine Möglichkeit, sensible Frage zu erheben, ist die *random response technique* (Warner, 1965). Dabei ist die Beantwortung einzelner Fragen mit einem Zufallsmechanismus verbunden. Das wäre ein Beispiel für einen Fragetyp, der sicher nicht als Kovariate geeignet ist!

(2002); Särndal und Lundström (2005); Bosnjak (2002) und Tourangeau et al. (2000).

Einige Aspekte, für die gezeigt werden konnte, dass sie das Responseverhalten beeinflussen, konnten nicht berücksichtigt werden, da sie sich mit der jeweiligen Befragung ändern oder auf andere Art und Weise nicht adäquat sind. Eine Liste der wichtigsten nicht berücksichtigten Faktoren findet sich in Abschnitt 4.5, ab S. 81.

4.1.3 Demografische Variablen

Dass es einen Zusammenhang zwischen soziodemografischen Merkmalen der potentiell Befragten, wie deren Alter und sozialem Status, gibt, konnte schon häufig gezeigt werden (Erbslöh und Koch, 1988; Reuband und Blasius, 2000). Allerdings handelt es sich bei diesen Merkmalen nur um «Globalvariablen» (Schnell, 1997), d. h. sie sind nicht ursächlich für die Teilnahmebereitschaft, sondern selbst wieder nur mit den zugrunde liegenden Charakteristiken korreliert. Die implizit angenommenen Mechanismen sind kaum untersucht (Schnauber und Daschmann, 2008).

Da Untersuchungen der Nutzungshäufigkeit des Internets immer auf sozio-demografischen Angaben beruhen (Froidevaux und Täube, 2006), sollen sie hier trotzdem auch berücksichtigt werden. Untersuchungen über weitere Unterschiede in den Charakteristika von Internetnutzern und solchen Personen, die das Internet nicht so häufig nutzen, existieren meines Wissens nicht. Auch deswegen wurden demographische Angaben schon erfolgreich zur Schätzung von *propensity scores* eingesetzt (Ekholm und Laaksonen, 1991).

Obwohl der Informationsgehalt demographischer Variablen zur Schätzung der Wahrscheinlichkeit, an einer Online-Befragung teilzunehmen, nicht ausreicht, handelt es sich doch um wichtige Informationen. Zum Beispiel konnte gezeigt werden, dass bei Online-Befragungen junge, hochgebildete Männer überrepräsentiert sind
(Schoen und Faas, 2005). Es gibt allerdings mindestens im Offline-Bereich Befragungen, bei denen sich kaum demografische Unterschiede zwischen Teilnehmenden und nicht Teilnehmenden zeigen, die aber trotzdem einen Bias aufweisen (Sidenvall et al., 2002).

Erste Kandidaten für Variablen zur Schätzung der *propensity scores*

sind daher demografischer Natur[7]. Der Fragebogen, wie er auch in der experimentellen Befragung verwendet wurde, befindet sich im Anhang, Abschnitt C, ab S. 241ff. Die Nummer der Frage im tatsächlichen Fragebogen steht jeweils in eckigen Klammern hinter der Frage[8].

1. Alter [1]

2. Geschlecht [2]

Natürlich korrelieren sozio-demografische Variablen noch mit sehr vielen anderen Charakteristiken, die einen Einfluss auf die Kooperationsbereitschaft haben. Zum Beispiel diskutieren Groves und Couper (1998, S. 120) hohes Alter als Variable, die sowohl einen positiven wie auch negativen Einfluss auf die Teilnahmebereitschaft haben kann. Einerseits unterstellen sie älteren Menschen ein höheres Mass an Bürgerpflicht, was sich positiv auf die Teilnahmebereitschaft auswirkt, andererseits führt eine undifferenzierte Angst vor Verbrechen bei älteren Menschen zu einer ablehnenden Haltung.

Zudem haben persönliche Variablen Einfluss auf die Wahrscheinlichkeit, erreicht zu werden. Diese Variablen, zu denen z. B. die Arbeitsbelastung gehört, korrelieren mit der sozialen Schicht der zu befragenden Person. Bei vielen Befragungen kommt es daher zu einem Mittelschichtbias, d. h. dass Personen aus den oberen und unteren sozialen Schichten unterrepräsentiert sind (Hartmann und Schimpel-Neimanns, 1992a).

Sozio-ökonomischer Status

Der sozio-ökonomische Status kann mithilfe verschiedener Skalen gemessen werden. Für einen Überblick siehe Wolf (1995). Allerdings beruhen die dort vorgestellten Skalen alle auf einer umfangreichen Fragenbatterie. Dies soll hier vermieden werden, da die Fragen nur der Gewichtung dienen sollen und keinen zu grossen Raum in der Befragung einnehmen dürfen.

7 Im Folgenden werden nicht die tatsächlichen Formulierungen der Fragen aus dem Fragebogen der experimentellen Befragung übernommen. Es handelt sich bei der Befragung um eine CATI-Befragung, daher ist die Formulierung oft etwas ungewohnt, zumal das die Befragung durchführende Marktforschungsinstitut LINK auf eine Anpassung der Formulierungen an die Schweizer Mundart bestanden hat. Ausserdem ergeben sich einige Angaben, wie z. B. dem Geschlecht der Befragten aus der Befragung selbst und werden nicht erfragt. Zu Details der Befragung siehe Kapitel 5, S. 87ff.

8 Die endgültige Reihenfolge der Fragen wurde von LINK bestimmt, was auch dazu geführt hat, dass Fragen zu einem thematischen Schwerpunkt, wie eben z. B. der Demografie an unterschiedlichen Stellen im Fragebogen gestellt werden und nicht «en bloc». Die grosse Erfahrung von LINK war bei der Umsetzung der Befragung sehr hilfreich.

Eine kurze Alternative zu den traditionellen Skalen ist eine Selbsteinschätzungsfrage, wie sie regelmässig in der Marktforschung verwendet wird, z. B. in der Allensbacher Markt- und Werbeträgeranalyse[9]. Allerdings zeigt die Erfahrung, dass derartige Fragen zu vielen Verweigerungen führen und sie sehr inkonsistent beantwortet werden. Deswegen soll hier die soziale Schicht mittels Ausbildung und Einkommen grob geschätzt werden.

3. Ausbildung [47]

4. Einkommen [48, 49]

Daneben können noch weitere demographische Charakteristika erfragt werden, die sowieso standardmässig durch LINK erhoben werden und hier genutzt werden können.

5. Zivilstand [45]

6. Zusammensetzung des Haushalts [46]

7. Wohnort (und damit auch Sprach- und Grossregion) [50]

Auch diese Variablen sind plausible Kandidaten für das PSA. Zivilstand und insbesondere die Zusammensetzung des Haushalts sind Indikatoren für die Zeit, die potentielle Befragte für Befragungen zur Verfügung haben. Personen aus Haushalten mit kleinen Kindern haben vermutlich weniger Zeit, sich an Befragungen zu beteiligen. (Empirische Evidenz für diese Vermutung konnte nicht gefunden werden. Ein gegenläufiger Effekt könnte sein, dass solche Personen häufiger zu Hause anzutreffen sind.) Dass sich die Teilnahmebereitschaft regional unterscheidet, ist ein bekanntes Phänomen. So ist beispielsweise die Teilnahmebereitschaft im Tessin geringer als in den anderen Landesteilen.

4.2 Rational Choice

Schnell (1997) und Groves und Couper (1998) fassen die Entscheidung für

9 www.awa-online.de

oder gegen die Teilnahme als rationale Wahl auf[10]. Die potentiellen Befragten treffen im Moment der Interviewanfrage die Entscheidung zu partizipieren oder nicht zu partizipieren (bei *self administered* Befragungen kommt noch die Option «Aufschub» hinzu). Die Entscheidung treffen die angefragten Personen aufgrund einer Kosten-Nutzen-Abwägung. Sie wählen diejenige Option, die den höchsten Nutzen, d. h. die grösste Bedürfnisbefriedigung verspricht. Diese Entscheidung ist (vermutlich) weniger eine bewusste Kalkulation, sondern läuft heuristisch ab (Groves und Couper, 1998; Schnauber und Daschmann, 2008).

3 Annahmen

Gemäss Bamberg et al. (2008) lassen sich drei Annahmen als nomologischer Kern der Theorie Rationaler Wahl (*rational chocie theory, (RCT)* postulieren:

Die Präferenz-Annahme Individuelle Präferenzen (oder Ziele) sind die Determinanten von Handlungen. Handlungen dienen instrumentell der Befriedigung dieser Präferenzen.

Die Restriktions-Annahme Alles, was die Fähigkeit eines Individuums erhöht oder verringert (d.h. Gelegenheiten oder Restriktionen), durch die Ausführung bestimmter Handlungen seine Präferenzen zu befriedigen, determiniert die Ausführung dieser Handlungen.

Die Nutzenmaximierungs-Annahme Individuen wählen diejenige Handlungsoption aus, mit der sie unter den gegebenen Restriktionen ihre Präferenzen am besten befriedigen.

Die eigentliche Schwierigkeit in der Anwendung der RC-Theorie liegt darin, geeignete Hilfs- oder Brückenannahmen zu finden, die es ermöglichen, diesen nomologischen Kern auf ein konkretes Problem anzuwenden. Sehr schnell kann es zu einer Tautologiesierung kommen, da sich ex post alles als Nutzen oder Kosten interpretieren lässt (Kirchgässner, 2008). Die praktische Schwierigkeit besteht darin festzulegen, welche Dinge als relevante Nutzen und Kosten betrachtet werden können und welche nicht.

Enge und weite RCT

Innerhalb der RC-Theorie haben sich grob vereinfacht zwei Standpunkte

10 Die Theorie Rationalen Handelns (Rational Choice Theory, RC Theory) ist eine Sammelbezeichnung für verschiedene Handlungstheorien. Entsprechend existiert keine einheitliche Definition bzw. kein einheitlicher Rahmen. Als Überblick und Einstieg dient neben vielen anderen Kirchgässner (2008).

entwickelt. Es gibt Vertreter einer *engen* Variante der RCT, wie sie beispielsweise Esser (1996) umreisst, und Vertreter einer *weiten* Variante. Sie unterscheiden sich dadurch, dass erstere nur «harte» Restriktionen akzeptieren[11]. «Hart» bedeutet, dass nur Belohnungen und Bestrafungen relevant sind. Hier soll die weite Variante der RCT vertreten werden, da sie sich bei vergleichbaren sozialen Situationen als Theorie mit hoher Erklärungskraft erwiesen hat.

Teilnahme ist low-cost Situation

Bei der Erklärung von Teilnahmeverhalten handelt es sich im Sinn der RCT um eine so genannte Low-Cost-Situation, also um eine Situation, in der der Unterschied zwischen harten Anreizen und Kosten der Handlung sehr gering ist (Preisendörfer, 1999; Schnell, 1997). Für andere Beispiele von Low-Cost-Situationen siehe z. B.
Seipel und Eifler (2003) für die Kriminologie und Preisendörfer (1999) in der Umweltsoziologie. Bei diesen Low-Cost-Situationen hat die harte Variante der RCT Theorie, wie sie z. B. in der neo-klassischen Ökonomie vertreten wird[12], keine hohe Prognosekraft. Die Definition von Nutzen und Kosten kann im Kontext dieser Arbeit also eher grosszügig.

4.2.1 Kosten von Befragungen

Privatsphäre

Als grösster Posten auf Seite der Kosten steht das «Eindringen in die Privatsphäre». Bei einer Befragung durch Meyers und Oliver (1978) geben 12% der Verweigerer «Concerns of privacy» als Hauptgrund für die Verweigerung an; bei DeMaio (1983) sind es 17%.

Es ist nicht klar, ob Privatheit im Kontext von Online-Befragungen immer noch so einen hohen Stellenwert hat. Aktuellere als die genannten Studien mit dem Schwerpunkt auf dem Vergleich verschiedener Verweigerungsgründe konnten bisher leider nicht gefunden werden. Auch die neueren Abhandlungen wie z. B. von Schnell (1997) und Neller (2005) beziehen sich wiederum auf Meyers und Oliver (1978) und DeMaio (1983).

Einerseits lässt sich ein wachsendes Schutzbedürfnis in Bezug auf per-

11 Die Unterscheidung betrifft noch weitere, hier weniger wichtige Punkte. Für eine Diskussion siehe Bamberg et al. (2008).

12 Dieser Ansatz wird in den meisten Lehrbüchern der Ökonomie (im Sinne von Volkswirtschaftslehre) vertreten (als prominentes, immer noch aktuelles Beispiel z. B. Samuelson und Nordhaus (1948)), obwohl sich die Disziplin insgesamt schon weiterentwickelt hat.

sönliche Daten beobachten (Viseu et al., 2004). Andererseits existiert ein Trend, Privates über Medien zu verbreiten (*self-disclosure*) (Weil, 2005). Es ist nicht klar, inwiefern sich diese beiden Tendenzen zukünftig auf das Antwortverhalten auswirken werden.

Ein Indikator für das Schutzbedürfnis des Privaten ist schwierig zu finden. Reinecke (1991) schlägt eine Skala mit zehn Items vor, die daher für den Kontext dieser Befragung zu lang ist. Eines dieser Items aufgreifend soll versucht werden, das Schutzbedürfnis mit folgender Frage zu messen:

8. Ich bin bereit, über meine persönlichen Gewohnheiten auch mit jemandem zu sprechen, den ich nicht so gut kenne. [35]

Natürlich ist es dabei von besonderer Wichtigkeit zu wissen, inwiefern eine mögliche Aversion persönliche Dinge preiszugeben, auch bei Befragungen, insbesondere auch im kommerziellen Kontext, relevant ist:

9. Bei Umfragen wird häufig öppes[13] gefragt, was niemand etwas angeht. [34]

10. Marktforschungsunternehmen behandeln die Daten vertraulich. [32]

Katz und Tassone (1990) zeigen, dass der Anteil der Personen, die angeben, aus Sorge um die «privacy» verweigert zu haben, umso geringer ist, je mehr Gründe zur Verweigerung zur Auswahl gestellt werden. Es bleibt allerdings unklar, ob es sich dabei um einen gewöhnlichen Vorgabeeffekt handelt oder ob *privacy* tatsächlich – so wie es auch Schnell (1997) behauptet – als Verweigerungsgrund überschätzt wird. Eindeutige Ergebnisse liegen nicht vor (Singer et al., 1993).

Aufwand

Ein weiterer in den Befragungen genannter Grund zu verweigern, ist der Aufwand, den die Befragung verursacht. Als Indikatoren sollte folgende Frage dienen:

13 LINK formuliert Fragen, die via Telefon gestellt werden, immer leicht der Schweizer Umgangssprache angepasst. Die Hoffnung ist, so ein grösseres Vertrauen zu schaffen und den «schulprüfungsartigen» Charakter von in Schriftdeutsch gestellten Fragen zu vermeiden. Diese Eingriffe in die Frageformulierung fallen je nach Interviewer unterschiedlich aus. LINK bietet die Möglichkeit an, dass Auftraggeber bei telefonischen Interviews zuhören können. Der Kunde kann passiv an den Interviews teilnehmen, ohne dass die Interviewer dies merken. Dabei konnte festgestellt werden, dass manche Interviewer die Fragen spätestens dann, wenn der Interviewte zögerlich antwortet oder eine Antwort zu verweigern droht, relativ frei umformuliert. Das mag bezüglich der Validität nachteilig sein, hat aber den Vorteil, dass der Item-Nonresponse gesenkt wird.

- Stimmen Sie folgender Aussage zu: Ich empfinde es als aufwändig, Befragungen zu beantworten.

Allerdings war LINK nicht bereit, diese Frage zu stellen. Die Befürchtung, den Befragten nochmals vor Augen zu führen, dass Befragungen einen Aufwand darstellen und es deswegen zu zukünftigen Verweigerungen oder Abbrüchen kommt, war leider zu hoch.

Erfahrung mit dem Medium

El-Menouar und Blasius (2005) konnten zeigen, dass die Bereitschaft an Online-Befragungen teilzunehmen, auch mit der Interneterfahrung zusammenhängt. Es kann unterstellt werden, dass mit steigender Interneterfahrung der kognitive Aufwand sinkt, an einer Befragung im Internet teilzunehmen. Sie messen die Erfahrung mit der Frage: «Wie schätzen Sie Ihre Interneterfahrung ein?» Da die experimentelle Befragung im Rahmen der standardisierten Panel-Rekrutierungs-Befragung von LINK durchgeführt wurde, wurde die bei LINK übliche Frage zur Erfahrung verwendet:

11. Denken Sie jetzt einmal an die Nutzung vom Internet für private Zwecke, es spielt dabei keine Rolle, wo, d.h. von welchem Anschluss aus Sie das Internet nutzen. Wie häufig etwa nutzen Sie das Internet für private Zwecke? Ist das. . .
    - mehrmals täglich
    - einmal täglich
    - mehrmals pro Woche
    - einmal pro Woche
    - seltener
    - nie
    - weiss nicht

    [37]

12. Wie oft haben Sie in den letzten 12 Monaten etwas online im Internet gekauft? [38]

13. Haben Sie zu Hause einen Breitband-Internet-Anschluss mit ADSL oder Kabelmodem oder nicht? Bei einem Breitbandanschluss zahlen Sie eine fixe Monatsgebühr pro Monat für Ihren Internet-Anschluss. Sie können immer online sein, ohne Zusatzkosten zu haben. [39]

Die letzte Frage ist eine Standardfrage von LINK. Meiner Meinung nach ist sie etwas verwirrend, da «Breitband» bedeutet, eine hohe Datenübertragungsrate zu haben. Es gibt allerdings keine technische Definition von Breitbandanschluss bzw. «hohe Datenübertragungsrate».Ursprünglich wurde der Begriff verwendet, um Zugangswege zum Internet zu bezeichnen, die schneller als die Einwahl mittels Telefonmodem oder ISDN sind, was heute allgemeiner technischer Standard ist. Die von LINK gegebene Erklärung bezieht sich aber eher auf die Form des Abonnements und bezeichnet eher das, was man gemeinhin als Flatrate als besonderer Form des Internet-Abonnements kennt.

4.2.2 Nutzen von Befragungen

Einstellung gegenüber Befragungen

Positive Einstellungen gegenüber Befragungen gehören sicher mindestens im Sinne der weiten Rational Choice Theorie zum Nutzen einer Handlung (Bamberg et al., 2008). Stocké und Becker (2004) können zeigen, dass eine positive Einstellung Befragungen gegenüber die Teilnahmebereitschaft erhöht. Umfragen zum Umfrageverhalten zeigen, dass die Teilnehmer auch aus Interesse und Spass an Befragungen teilnehmen (Schleifer, 1986; Stocké und Langfeldt, 2003b). Interesse betrifft dabei sowohl Selbsterkenntnis als auch Interesse am Thema.

Um den Spass an Befragungen bzw. die positiven Einstellungen ihnen gegenüber zu erfassen, wird die folgende Frage als Indikator verwendet:

14. Umfragen bringen Abwechslung und sind interessant. [33]

Incentives

Auch materielle Dinge können ein Anreiz sein, an Befragungen teilzunehmen. Es lässt sich empirisch zeigen, dass so genannte *Incentives* die Teilnahmebereitschaft erhöhen (Church, 1993; Porter und Whitcomb, 2003, ,. . .). Dies konnte auch für Online-Panel-Befragungen gezeigt werden (Göritz und Wolf, 2008). Diese Ergebnisse sind allerdings nicht unumstritten: Warriner et al. (1996) zeigt, dass nur Anreize, die ex ante ausgezahlt werden, eine positive Auswirkung auf die Teilnahmebereitschaft haben.

Die Befundlage zu materiellen Anreizen bei Befragungen ist etwas widersprüchlich und unübersichtlich (Bonke und Fallesen, 2010). Es ist zu vermuten, dass es grosse Unterschiede hinsichtlich des Kontextes der Befragungen gibt. So kann man sich vorstellen, dass die intrinsische Motivation

bei amtlichen und akademischen Befragungen höher ist, als bei kommerziellen.

Als Frage zum Thema wird in die experimentelle Befragung übernommen:

15. Dass es für die Teilnahme an Online-Befragungen eine Belohnung gibt, ist für mich ein wichtiger Grund zum Mitmachen. [43]

Ein negativer Effekt materieller Anreize ist —nebenbei bemerkt— das häufige Auftreten so genannter *lurkers*. Das sind Personen, die nur wegen der materiellen Anreize an der Befragung teilnehmen und sich nur völlig unreflektiert durch die Befragung klicken. Siehe dazu z. B. Bosnjak (2002) und für Methoden, resultierende Fehler im Kontext von Panels zu identifizieren, Schroepler und Wagner (2005).

Zum Thema Incentives gibt es eine umfangreiche Literatur, die die verschiedenen Möglichkeiten und Variationen, Anreize zu setzten, vergleicht. Diese soll hier nicht besprochen werden; als Einstieg in das Thema sei neben Göritz und Wolf (2008) auch auf Su et al. (2008) für den Bereich Online-Befragungen verwiesen; für allgemeine Befragungen bieten Groves et al. (2004) einen guten Überblick.

Bei der experimentellen Befragung sind Tests zu verschiedenen Anreizformen weder vorgesehen noch möglich. Insbesondere lässt sich die Art oder der Zeitpunkt des Anreizes nicht variieren, wie beispielsweise der Möglichkeit, die Incentives vor der Befragung auszuzahlen, unabhängig von der effektiven Teilnahme.

LINK bietet verschiedene Arten der Incentivierung an: neben COOP und Migros Kundentreuepunkten (Supercard-Punkte bzw. Cumulus-Punkte), wechselnden unmittelbaren Belohnungen (Piccolo-Flasche Sekt) können Befragte auch die Option wählen, dass ein Geld-Betrag einer wiederum wählbaren wohltätigen Organisation überwiesen wird. Insbesondere die letzte Option ist bei den LINK Panelisten sehr beliebt, in anderen Ländern haben Panel-Betreiber damit weniger bis keinen Erfolg (für Deutschland: Drewes (2010)).

Langfristig könnte die Wahl der Anreize auch Teil des PSA werden.

4.3 Persönlichkeitsmerkmale

Der Einfluss von Persönlichkeits- und Charaktereigenschaften ist in den Sozialwissenschaften lange ignoriert worden (Borghans et al., 2008). Erst allmählich hat sich in den letzten Jahren die Erkenntnis durchgesetzt, dass sie ein interessanter Analysegegenstand sein können. Das gilt sowohl für Persönlichkeitsvariablen als abhängige wie auch als erklärende Variable, wie z. B. bei der Untersuchung sozialer Ungleichheit (Roberts et al., 2004).

Offensichtlich beeinflussen solche Persönlichkeitseigenschaften auch das Antwortverhalten (Neller, 2005; Schnauber und Daschmann, 2008). Es konnte gezeigt werden, dass nicht eigentlich die viel untersuchten demografischen Angaben die entscheidenden Einflussfaktoren für die Teilnahmebereitschaft sind, sondern dass die zugrunde liegenden Mechanismen Persönlichkeitseigenschaften wie soziale Isolation, generelles Misstrauen oder Introvertiertheit verantwortlich sind (Schnauber und Daschmann, 2008; Esser, 1973; Goyder, 1987). Diese Eigenschaften sind bisher im Kontext der Schätzung von Teilnahmebereitschaft noch wenig untersucht.

Die differentielle Psychologie ist der Teil der Psychologie, der sich sich mit der Erfassung und Messung von Unterschieden in Persönlichkeitseigenschaften beschäftigt. Es wurden verschiedene Instrumente entwickelt, um die verschiedenen Aspekte der Persönlichkeitsstruktur zu messen. Für einen Überblick siehe z. B. die Lehrbücher von Friedman und Schustack (2005); Larsen und Buss (2007); Reis (2000) und Robins und Fraley (2007). Einige dieser Instrumente werden auch in den grossen, repräsentativen Befragungen eingesetzt (Rammstedt und John, 2007).

Big Five

Ein bekanntes und einflussreiches Konzept, Persönlichkeitsunterschiede zu messen, ist der Big-Five-Ansatz[14]. Der Big-Five-Ansatz hat sich nicht nur in der (Sozial-) Psychologie etabliert, sondern in der gesamten Sozialwissenschaft. So verwendet z. B. das Sozio-Ökonomische Panel (SOEP) des Deutschen Instituts für Wirtschaftsforschung (DIW) seit 2005 diesen Ansatz, um die Persönlichkeit der Befragten zu charakterisieren, um dann mit diesen Persönlichkeitsunterschieden verschiedenste Verhaltensmuster zu erklären (Gerlitz und Schupp, 2005).

14 Für eine Übersicht über die Entwicklung des Big-Five Ansatzes siehe z. B. Urbina (2004) und Friedman und Schustack (2005)

4.3.1 Das Big-Five Inventory

Im Mittelpunkt des Ansatzes steht die Annahme, dass Unterschiede in der Motivation und im Handeln von Individuen auf Unterschiede in ihren Persönlichkeiten zurückführbar sind. Persönlichkeit wird dabei als fünfdimensional betrachtet. Die fünf Dimensionen sind Neurotizismus (N), Extraversion (E), Offenheit für Erfahrungen (O), Verträglichkeit (V) und Gewissenhaftigkeit (G). Diese Big-Five Dimensionen unterteilen sich dann wiederum in jeweils sechs Subdimensionen, so genannte Facetten. Details zum Big-Five-Ansatz finden sich eigentlich in jedem Lehrbuch zu Persönlichkeitstheorien. Die folgenden Angaben stammen aus Amelang (2006), Laux und Gessner (2008), Pervin et al. (2005) und Simon (2006).

Nach dem psycho-lexikalischen Ansatz in der Tradition George Allports (Allport, 1937) lässt sich die Persönlichkeit vollständig mit Adjektiven beschreiben. Allport wählte 35 Adjektive als essentiell aus. Tupes und Christal (1992) reduzierten diese dann später auf fünf und begründeten damit den Big-Five-Ansatz.

Dimensionen

Neurotizismus Die Neurotizismusskala bildet die Art und Weise ab, wie Emotionen erlebt werden und wie mit ihnen umgegangen wird. Je ängstlicher, nervöser, launischer, empfindlicher, reizbarer und furchtsamer ein Mensch ist, desto höher ist sein Neurotizismuswert. Hohe Neurotizismuswerte findet man bei Personen, die emotional labil sind.

Extraversion Diese Skala bildet den Grad ab, mit der eine Person mit der Umwelt interagiert. Extraversion ist ein Pool auf einer Skala deren anderes Ende die Intraversion (Introvertiertheit) bildet. Extravertierte Menschen sind gesprächig, bestimmt, aktiv, energisch, dominant, enthusiastisch und abenteuerlustig. Intravertierte dagegen still, sorgfältig, scheu, reflektierend und zurückgezogen.

Offenheit für Erfahrungen Diese Skala soll das interesse und Ausmass der Beschäftigung mit neuen Aktivitäten abbilden. Menschen mit viel Offenheit werden lexikalisch charakterisiert durch Adjektive wie

- einfallsreich, originell, erfinderisch, phantasievoll
- intellektuell neugierig, offen für neue Ideen

- interessiert an Ästhetischem wie Kunst, Musik und Poesie
- mit Vorliebe für Abwechslung (statt Routine), Neigung zu neuen Aktivitäten, neuen Reisezielen, neuem Essen usw.
- aufmerksam für eigene und fremde Emotionen
- bereit, traditionelle Werte in Frage zu stellen

Verträglichkeit Verträglichkeit ist neben Extraversion eine zweite Skala, die den Umgang mit anderen Menschen abbildet.

Menschen, die altruistisch und hilfsbereit sind, haben einen hohen Wert auf der Verträglichkeitsskala. Sie sind charakterisierbar durch Adjektive wie mitfühlend, nett, warm, vertrauensvoll, hilfsbereit, kooperativ und nachsichtig.

Gewissenhaftigkeit Gewissenhafte Menschen zeichnen sich aus durch Adjektive wie ordentlich, genau, organisiert, sorgfältig, verantwortlich, zuverlässig und überlegt.

Es gehört zu den Allgemeinplätzen der differentiellen Psychologie, dass Persönlichkeitsunterschiede konstant sind (Amelang, 2006). Sollte sich zeigen, dass Persönlichkeitsunterschiede ein guter Prediktor für das Teilnahmeverhalten sind, wären sie eine grosse Hilfe, gerade im Kontext von Panel-Befragungen.

4.3.2 Messung der Big-Five

Es gibt verschiedene Skalen, um einen Big-Five Index zu bestimmen. Für das SOEP wurden viele dieser Skalen getestet (Gerlitz und Schupp, 2005; Dehne und Schupp, 2007).

McCrae und Costa (1985) haben als Instrument das NEO-Personality Inventory (NEO-PI) zur Messung der Big-Five vorgelegt und später zum NEO-Personality Inventory Revised (NEO-PI-R) erweitert. Es handelt sich dabei um eine Batterie von 240 Items, deren Beantwortung rund 45 Minuten dauert. Parallel heben sie eine Kurzskala, das Neo-Five-Factory-Inventory (NEO-FFI) vorgestellt, bei der die Beantwortung der Fragen rund 15 Minuten

benötigt[15].

Es wurde eine Reihe weiterer Kurzskalen entwickelt, um die Messung der Big-Five auch im Rahmen von Befragungen, die nicht den Schwerpunkt auf die Messung der Big-Five legen, zu ermöglichen[16]. Es konnte in verschiedenen Studien gezeigt werden, dass die Charakterisierung mittels Kurzskalen denen der längeren Skalen nahe kommen und daher ein akzeptabler Ersatz für diese sind. Burisch (1997) kommt zu dem allerdings nur theoretisch abgeleiteten Ergebnis, dass kurze Skalen ähnlich gut die Persönlichkeitsunterscheide messen können wie lange Skalen.

Lang et al. (2001) vergleichen das von John et al. (1991) entwickelte Big-Five-Iventory (BFI-44) mit 44 Items mit dem NEO-PI-R. Sie kommen zu dem Ergebnis, dass die psychometrischen Eigenschaften des BFI-44 äquivalent zu denen des NEO-PI-R sind. Rammstedt und John (2007) wiederum vergleichen eine Version des BFI-10 mit zehn Items mit dem BFI-44 und dem NEO-PR-I jeweils in der englischen und deutschen Version. Der BFI-10 hat wiederum ähnliche psychometrische Eigenschaften wie der BFI-44, ist aber schon weiter als dieser vom NEO-PR-I entfernt. Ausserdem konnten Arbeiten auf Grundlage von Paneluntersuchungen zeigen, dass der BFI-10 im Zeitverlauf weniger konstant ist als der BFI-44. Trotzdem kommen Rammstedt und John zu dem Fazit, dass die Informationsverluste durch den Gewinn an Kürze gerechtfertigt sind.

Gosling et al. (2003) entwickelten einen dem BFI-10 ähnlichen Index mit modifizierten Fragen, dem Ten Item Personality Inventory (TIPI). Es zeigt sich, dass die Korrelation zwischen TIPI und BFI-44 etwas höher ausfällt, als zwischen BFI-10 und BFI-44. Gerlitz und Schupp (2005) testen TIPI, BFI-44 und BFI-10 im Kontext des SOEP. Sie kommen einerseits zu dem Resultat, dass das BFI-44 zu lang ist, d. h. dass die Befragten zu viel Zeit benötigen zum die Fragenbatterie zu beantworten. Andererseits bildet das BFI-10 nicht ausreichend reliabel das NEO-PI-R ab. Als Alternative leiten Gerlitz und Schupp aus dem BFI-44 eine Skala ab, die 25 Items umschliesst.

15 Die Übersetzung ins Deutsche des NEO-PI-R erfolgte durch Angleitner und Ostendorf (2004) und des NEO-FFI durch Borkenau und Ostendorf (1993).

16 Für eine Übersicht siehe z. B. Urbina (2004). Für Instrumente, die für den deutschen Sprachraum entwickelt wurden, Gerlitz und Schupp (2005).

4.3.3 Ableitung der Skala

Man kann vermuten, dass nicht alle fünf Persönlichkeitsdimensionen des Big-Five einen entscheidenden Einfluss auf die Teilnahmebereitschaft an Online-Befragungen haben. Es wird notwendig sein, empirisch zu überprüfen, welchen Dimensionen ein grösseres Gewicht zukommt.

Generell ist es so, dass die Persönlichkeit mit zunehmender Skalenlänge besser erfasst und beschrieben werden kann. Und umso besser dies funktioniert, desto genauer sollte auch die Teilnahmebereitschaft prognostiziert werden können. Da die Anzahl der Fragen, die in der experimentellen Befragung gestellt werden können, sehr beschränkt ist, ist es vertretbar, eine kürzere Skala zu verwenden[17].

BFI-S

Bei der experimentellen Befragung soll das BFI-S verwendet werden. Das BFI-S ist eine von Gerlitz und Schupp (2005) entwickelte 15 Items umfassende Kurzskala, abgeleitet aus dem TIPI. Die BFI-S wird in der gleichen Formulierung auch im SOEP verwendet. Die Beantwortung des BFI-S dauert nur rund 2 Minuten. Hinsichtlich Reliabilität und Validität schneidet das BFI-S verglichen mit dem NEO-PI-RR zufriedenstellend ab (Gerlitz und Schupp, 2005; Dehne und Schupp, 2007). Das bedeutet, dass die interne Konsistenz der Skalen hinreichend gut ist und die Trennschärfe zwischen den Items hoch (Dehne und Schupp, 2007).

Tabelle 4.1 listet die aus der Big-Five Skala übernommenen Fragen mit ihrer jeweiligen Position im Fragebogen auf. Die Dimensionen sind **N** Neurotizismus, **E** Extraversion, **O** Offenheit für Erfahrungen, **V** Verträglichkeit und **G** Gewissenhaftigkeit.

Die genaue Frageformulierung lautet: « Die folgenden Aussagen beschreiben Eigenschaften und Einstellungen, wo auf einen mehr oder weniger zutreffen können. Bitte sagen Sie, inwiefern diese Aussagen Ihrer Meinung nach auf Sie zutreffen.

Sie können Ihre Antwort jeweils zwischen 1 und 5 abstufen. Dabei bedeutet 1 ‹trifft auf mich überhaupt nicht zu› und 5 bedeutet ‹trifft auf mich voll und ganz zu›.»

Die genaue Reihenfolge der Items wurde zufällig bestimmt, die Nummerie-

17 Das Ziel dieser Arbeit ist es ja nicht, die Persönlichkeit von Befragten im Vergleich zu Verweigerern präzise zu beschreiben, sondern möglichst gute Prädiktoren für das Teilnahmeverhalten zu finden.

rung dient nur der Referenzierung in der Dokumentation. Die Formulierung der Items wurde immer so vorgenommen, dass sie immer mit «Ich bin jemand, der. . .» beginnt.

Bei der Skala wurden nur die Endpunkte benannt, es handelt sich also sozusagen um eine endpunktbeschriftete Skala, wie sie z. B. auch Porst (2008) empfiehlt.

Gerlitz und Schupp (2005) schlagen, pro Item eine siebener Rating-Skala vor, von «trifft überhaupt nicht zu» bis «trifft voll zu». Die Breite der Skala wurde dem LINK-Standard entsprechend auf eine fünfer Skala reduziert.

	Item	Dimension	
Ich bin jemand, der. . .	gründlich schafft	G	[7]
	kommunikativ und gesprächig ist	G	[8]
	manchmal ä chli grob zu anderen ist	E	[9]
	originell ist und Ideen einbringt	O	[10]
	sich oft Sorgen macht	N	[11]
	verzeihen kann	V	[12]
	eher faul ist	G	[13]
	aus sich herausgehen kann und gesellig ist	E	[14]
	künstlerische Erfahrungen schätzt	O	[15]
	leicht nervös wird	N	[16]
	Aufgaben wirksam und effizient erledigt	G	[17]
	zurückhaltend ist	E	[18]
	rücksichtsvoll und freundlich mit anderen umgeht	V	[19]
	eine lebhafte Phantasie hat	O	[20]
	entspannt ist und mit Stress gut umgehen kann	N	[21]

Tabelle 4.1: Fragen abgeleitet aus dem Big-Five Inventory

4.4 Werte

Definition

Werte werden definiert als wünschenswerte, situationsübergreifende Ziele, unterschiedlicher Wichtigkeit (Davidov et al., 2008). Sie haben in der Psychologie und in den Sozialwissenschaften eine lange Tradition, menschliches Verhalten zu erklären (Allport, 1937; Allport et al., 1970; Rokeach, 1973; Schwartz, 1992),
(Verplanken und Holland, 2002, ...)[18]. Auch in den Wirtschaftswissenschaften, insbesondere im Bereich der experimentellen Spieltheorie, werden Werte zunehmend als Motivator für menschliches Handeln akzeptiert, siehe z. B. Fehr und Gintis (2007) als aktuelle Übersicht[19].

Schwartz-Skala

In der empirischen Sozialforschung fristen Werte allerdings ein Schattendasein, da es kaum allgemein akzeptierte Skalen gibt, diese zu messen. Eine mögliche Skala schlägt Schwartz in Anlehnung und Erweiterung einer Skala von Rokeach vor (Schwartz, 1992; Rokeach, 1973). Schwartz erweitert die Skala später nochmals
(Schwartz et al., 2001). Die Skala ist auf breite Akzeptanz gestossen und ist insbesondere im Bereich interkultureller Vergleiche auf grosse Resonanz gestossen und hat zu einen «Boom» empirischer Studien geführt (Davidov et al., 2008). Dies gilt auch für den Bereich *intra*kultureller Studien, wie z. B. bei der Untersuchung politischer Einstellungen
(Feldman, 2003). Beispielsweise wurde die Skala in einer Pilotstudie des *American National Elections Surveys* verwendet[20].

Prominent wurde die Schwartz-Skala auch durch ihre Integration in das European Social Survey (ESS) (Stoop et al., 2002). Sie wurde in diesem Zusammenhang umfangreich getestet und aus dem Englischen in die hier relevanten Sprachen Deutsch und Französisch übersetzt (Davidov, 2002).

Schwartz unterscheidet zehn Gruppen motivationaler Werte, die hier nur

18 Für eine Übersicht über die Geschichte der Werte als soziologisches Forschungsthema siehe Spates (1983)

19 Einschränkend ist zu sagen, dass nicht alle denkbaren Werte akzeptiert werden, sondern diese nur sehr vorsichtig eingeführt werden, beispielsweise um Reziprozität zu erklären. Der Grund ist, dass einer Tautologisierung des Ansatzes vorgebeugt werden soll. Für eine Übersicht siehe z. B. Camerer (2003), Chaudhuri (2008) oder Gintis (2009).

20 Weitere Beispiele für die Verwendung der Schwartz-Skala finden sich bei Davidov et al. (2008)

stichpunktartig umschrieben werden sollen[21].

Macht Sozialer Status und Prestige, Kontrolle über Personen und Ressourcen

Leistung Persönlicher Erfolg durch Demonstration sozial akzeptierter und erwünschter Kompetenzen

Hedonismus Spass, sinnliche Belohnungen für sich selbst

Stimulation Reize, Neuheiten, Herausforderungen

Selbstbestimmtheit Gedankliche Unabhängigkeit, Freiheit, Entdecken können

Universalismus Verstehen, Wertschätzung, Toleranz und Schutz des Wohlbefindens anderer und der Natur

Benevolenz Schutz und Förderung der Wohlfahrt der Personen der unmittelbaren sozialen Umgebung

Tradition Respekt, Akzeptanz und Verpflichtung den Werten der eigenen Kultur und Religion gegenüber

Konformität Einschränkung von Handlungen, Neigungen und Impulsen die andere, oder soziale Erwartungen bzw. Normen verletzten könnten

Sicherheit Geborgenheit, Harmonie und Stabilität der Gesellschaft und sozialer Beziehungen

Ohne hier in grösserer Ausführlichkeit darauf eingehen zu wollen, unterstellt
Schwartz, dass die einzelnen Gruppen von Werten in Beziehung zueinander stehen. Die Beziehung ist ein «zirkuläres Kontinuum». Das bedeutet, dass sie sich ähnlich einem Farbkreis anordnen lassen können, es also sehr verwandte Werte gibt und solche, die sich gegenüberstehen. Die genaue Struktur dieses Wertekreises ist hier unerheblich, kann aber z. B. bei Davidov et al. (2008) nachgelesen werden.

21 Die genauen englischen Bezeichnungen lauten: *power, achivement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity* und *security.*

Aus diesen zehn Werten leitet Schwartz Frage-Items ab. In der ursprünglichen Version waren dies 40 Items, die aber schon im ESS in reduzierter Zahl (21) zur Anwendung gekommen sind (Stoop et al., 2002). Die Liste der Items aus dem ESS findet sich im Anhang, Kap. B auf S. 237.

Leider war es hier für die experimentelle Befragung nicht möglich, die Schwartz-Skala in der gekürzten ESS Variante zu übernehmen. Zwei wichtige Gründe waren ausschlaggebend. Zunächst ist die Skala zu lang. LINK konnte im Rahmen des Projekts nur eine maximal 12 Minuten dauernde Befragung durchführen.

Reduktion der Schwartz Skala

Es wurde versucht, solche Items auszuschliessen, die hoch mit einzelnen anderen Items korrelieren oder gesamthaft für eine hohe Multikollinearität sorgen. Dazu wurden die Daten des European Social Surveys (ESS) 2006 für die Schweiz verwendet[22]. Abbildung 4.1 zeigt einen Korrelationsplot aller Items der Schwarz-Batterie. Die Labels der Variablen sind wiederum im Anhang in Kap. B dokumwntiert.

Eine hohe Korrelation zwischen zwei Variablen bedeutet, dass die eine Variable mit Hilfe der anderen erklärt werden kann. Der Informationsverlust, den man durch eine Reduktion der Schwartz-Batterie in Kauf nimmt, ist umso kleiner, je besser die ausgeschlossene Variable durch eine andere Variable erklärt werden kann. Es ist daher weniger verlustreich, solche Variablen auszuschliessen, die hoch mit anderen Variablen korrelieren[23].

Ein zweites Auswahlkriterium war die inhaltliche Plausibilität der Items. Dazu wurden auch Gespräche mit Psychologen geführt. [24].

Es wurden zehn Items ausgewählt und für die experimentelle Befragung übernommen. Den reduzierten Korrelationsplot zeigt Abb. 4.2. Zu den Labels siehe auch Tabelle 4.2.

22 Sowohl der vollständige Fragebogen wie auch alle Daten sind auf der Web-Präsenz des ESS nach einer kostenlosen Anmeldung erhältlich: `http://ess.nsd.uib.no/`

23 In diesem Zusammenhang weniger von Bedeutung ist die Multikollinearität aller Variablen. Eigentlich ist es so, dass man auf eine Variable auch eher verzichten kann, je besser sie durch die anderen Variablen erklärt wird. Da es aber im Prozess der Modellierung zu einer weiteren Reduktion der Variablen kommen wird, ist die Multikollinearität zu diesem Zeitpunkt wenig aussagekräftig. Trotzdem nochmals illustrativ in Zahlen: die Multikollinearität konnte von $\kappa = 12.9$ auf $\kappa = 8.9$ reduziert werden.

24 Ich danke Dr. Anja Mücke für ihre Hilfe.

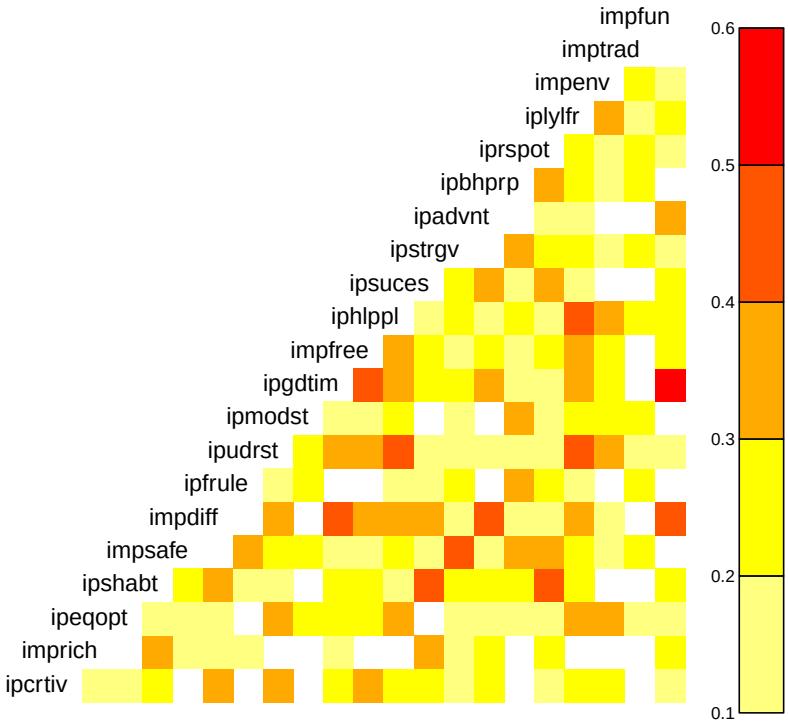


Abbildung 4.1: Korrelationsplot der Items der Schwartz-Skala wie im ESS mit den Daten des ESS

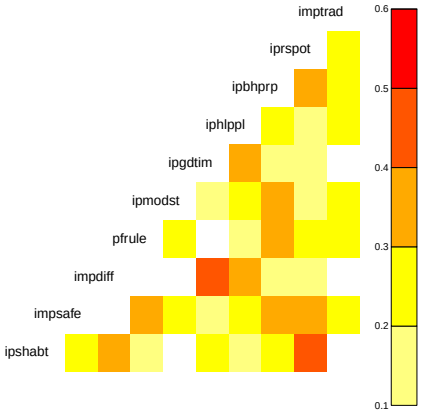


Abbildung 4.2: Korrelationsplot der Items der reduzierten Schwartz-Skala mit den Daten

Weitere Modifikationen

Eine weitere Restriktion bezüglich der Adaption der Schwartz-Skala aus dem ESS ergibt sich aus den unterschiedlichen Befragungsmedien. Das ESS ist eine face-to-face Befragung, was sehr viel längere Fragen erlaubt (Porst, 2008). Um die Fragen für eine CATI-Befragung zu adaptieren, musste die Itemlänge gekürzt werden. Ausserdem haben es die CATI-Experten von LINK für nicht möglich gehalten, den projektiven Fragestil des face-to-face Interviews zu übernehmen.

Dass die Fragen nicht dem genauen Wortlaut und nicht einmal dem Stil der gut getesteten Schwartz-Skala entsprechen, wird hier als unproblematisch angenommen. Es geht in diesem Kontext weniger darum, tatsächlich die individuellen Werte oder Persönlichkeitsmerkmale objektiv zu messen, als vielmehrdarum, gute Kovariate zu erheben. Je einfacher und leichter die Fragen für die Probanden zu beantworten sind, desto besser. Selbst wenn die Selbsteinschätzungsfragen, so wie sie hier formuliert werden, noch weniger objektiv die tatsächlichen Wertevorstellungen und Persönlichkeitseigenschaften messen, ist dies nicht so relevant, denn es geht ja nicht um deren tatsächliche Messung.

In Tabelle 4.2 sind die abgeleiteten Fragen für die experimentelle Befragung aufgeführt, jeweils versehen mit dem Verweis zum ursprünglichen Item aus der Schwarz-Skala.

«Auch die folgenden Aussagen beschreiben Eigenschaften und Einstellungen, wo auf einen mehr oder weniger zutreffen können. Bitte sagen Sie, inwiefern diese Aussagen Ihrer Meinung nach auf Sie zutreffen. Sie können Ihre Antwort jeweils zwischen 1 und 5 abstufen. Dabei bedeutet 1 ‹trifft auf mich überhaupt nicht zu› und 5 bedeutet ‹trifft auf mich voll und ganz zu›.»

Wie auch bei den Fragen aus dem Big-Five Inventory (Tabelle 4.1) ist die Position der Items in der experimentellen Befragung zufällig bestimmt worden. Beim Vorlesen wurde jedem Item die Formulierung «Ich bin jemand, dem es wichtig ist,. . .» vorangestellt. Die Persönlichkeitsmerkmale wurden also nicht mehr wie bei der Originalskala auf eine fiktive andere Person projeziert, sondern es handelt sich eher um (plumpere) Selbsteinschätzungen.

Ich bin jemand, dem es wichtig ist,...			
ipshabt	...	seine Fähigkeiten zu zeigen	[22]
impsafe	...	in einer sicheren Umgebung zu leben	[23]
impdif	...	immer wieder neue Aktivitäten auszuprobieren	[24]
imptrad	...	im Leben Abwechslung zu haben	[25]
impfrule	...	sich immer an Regeln zu halten, selbst dann, wenn es niemand sieht	[26]
ipmodest	...	zurückhaltend und bescheiden zu sein	[27]
ipgtim	...	Spass zu haben	[28]
iphlppl	...	den Menschen um mir herum zu helfen	[29]
ipbhprp	...	sich immer richtig zu verhalten und nicht den Eindruck zu erwecken, mich falsch zu verhalten	[30]
iprspot	...	von anderen respektiert zu werden	[31]

Tabelle 4.2: Abgeleitet Fragen aus der Schwartz-Skala

4.5 Nicht berücksichtige Einflussfaktoren

Da die Testbefragung begrenzt war, konnten nicht alle plausiblen bzw. alle in der Literatur bereits besprochenen Einflussfaktoren auf das Teilnahmeverhalten in die experimentelle Befragung übernommen werden. Ausserdem gibt es eine Reihe von aus der Literatur bekannten Einflussfaktoren, die sich für das PSA nicht eignen, da sie z. B. nicht als fixe Frage formuliert werden können. Im Folgenden werden die wichtigsten der nicht übernommenen Einflussfaktoren kurz umrissen.

Fragebogengestaltung Die Gestaltung des Fragebogens kann einen Einfluss auf die Teilnahmewahrscheinlichkeit haben (Dillman et al., 1998). Insbesondere gilt dies auch bei Online-Befragungen (Manfreda und Vehovar, 2002b).

Kommunikation mit Befragten Neben der Ankündigung einer Befragung (Dillman, 2000) kann auch eine Erinnerungsmail die Rate der Antwortenden erhöhen (Archer, 2008).

Allerdings gibt es widersprüchliche Befunde zur Bedeutung von Einladungen. Hembroff et al. (2005) findet einen positiven Zusammenhang

zwischen Einladungsschreiben und Teilnahmebereitschaft, Singer et al. (2000) dagegen nicht.

Thema Das Thema einer Befragung kann die Teilnahmebereitschaft beeinflussen (Tourangeau et al., 2000). Je mehr sich Personen für ein Thema interessieren, umso höher ist ihre Bereitschaft, sich an der Befragung zu beteiligen. Ausserdem hängt die Teilnahmebereitschaft vom Kontext der Befragung ab.

Besonderheiten der Erreichbarkeit Gelegentlich ist es so, dass es schwierig ist, die Person zu erreichen, die man tatsächlich erreichen möchte. Beispielsweise zeigen Petrie et al. (1997), dass es bei Befragungen von Unternehmen häufig einen mehrstufigen Ausfallprozess gibt, bevor die eigentlich zu befragende Person erreicht werden kann. Es existieren häufig Intermediäre, sogenannte *gatekeeper*, wie z. B. Sekretärinnen usw. Existieren solche Gatekeeper, ist es die Schätzung der *propensity scores* nicht mehr möglich. Von der Existenz solcher Gatekeeper soll daher im Folgenden abstrahiert werden.

Schwierigkeit Je schwieriger die Fragen sind, umso geringer ist die Antwortrate bei einzelnen Fragen (Tomaskovic-Devey et al., 1994).

Der Absender der Befragung Befragte nehmen auch aus altruistischen Motiven an Befragungen teil (Erbslöh und Koch, 1988). Dabei ist es so, dass die Hilfsbereitschaft höher ist, wenn der Absender der Befragung eine nicht kommerzielle Institution, wie z. B. eine Universität ist.

Befragungssituation Intuitiv ist klar, dass auch die Befragungssituation eine grosse Rolle dabei spielt, ob man sich an einer Befragung beteiligt. Ist man gerade im Stress oder wütend, wird man sich eher nicht an Befragungen beteiligen. Das Thema ist bisher nur sehr wenig untersucht (Groves und Couper, 1998; Schnauber und Daschmann, 2008). Es ist vorderhand völlig unklar, welchen Einfluss die Situation bei Online-Befragungen hat. Es ist nicht klar, wann welche Entscheidungsheuristik angewendet wird. Wenn die E-Mail mit der Einladung zu einer Befragung eintrifft, ist meist schon durch den Absender und die Betreffzeile erkennbar, um was es sich bei dieser E-Mail handelt. Es ist nicht beobachtbar, ob typischerweise in diesem Augenblick die

Entscheidung zur Partizipation oder Nicht-Partizipation fällt oder ob diese aufgeschoben wird. Es ist aber sehr wahrscheinlich, dass bei Befragungen über das Internet die Situation, in der die Befragten die Einladung erreicht, eine geringere Rolle spielt, da die Teilnahme aufgeschoben werden kann. Dies gilt für alle *self administered* Befragungen, nicht aber für telefonische und face-to-face Befragungen.

Schnauber und Daschmann (2008) unterscheiden bei den Einflussfaktoren der Teilnahmebereitschaft «Traits», also Charakteristiken der Befragten, die relativ konstant sind wie soziodemografische Merkmale, grundlegende Einstellungen und Einstellungen zu Umfragen von «States», worunter sie situatuionsabhängige Einflussfaktoren verstehen wie z. B. Merkmale der konkreten Befragung, situative Gegebenheiten und Interviewermerkmale. Sie können zeigen, dass es gerade die States sind, die die Teilnahmebereitschaft (bei CATI-Befragungen) beeinflussen.

Surveymanagement Das Management eines Web-Panels kann einen grossen Einfluss auf die gewonnen Daten haben. Die Qualität der Panelpflege ist, entscheidet, wie gross die Panelmortalität ist (Callegaro und Disogra, 2008). Panelpflege umfasst neben anderem Dinge wie der Häufigkeit, mit der mit den Panelmitgliedern kommuniziert wird, sie zu Befragungen eingeladen werden und wie gut die Verteilung der Incentives funktioniert. Um nur ein illustrierendes Beispiel zu nennen: Williams et al. (2006) können in einer Metastudie zeigen, dass die Responsequote kurzfristig enorm erhöht werden kann, indem man nur Befragte im Panel behält, die schon in der Vergangenheit sehr zuverlässig geantwortet haben. Selbstverständlich ist dadurch die aber Gefahr eines Bias sehr hoch.

Sonstiges Dillman (2000) schlägt das «tailored design» vor, womit weitere Eigenschaften von Befragungen gemeint sind, die das Rekrutierungsergebnis positiv beeinflussen können. Dies umfasst beispielsweise Dinge wie das Layout und das Wording.

Informationsstand und Wissen Weiterhin wurden das Wissen und der Informationsstand der Befragten zum Thema der Befragung nicht aufgenommen. Es lässt sich zeigen, dass mangelndes oder fehlerhaftes

Wissen zu instabilen Antworten führen kann (Wood, 1982) und die Bereitschaft zur Teilnahme reduziert (Kühne und Böhme, 2006). Die Erklärung für die Abhängigkeit der Teilnahmebereitschaft vom Stand des Wissens ist einerseits, dass das Wissen zu einem Thema umso höher ist, je höher die Zentralität, d. h. die persönliche Bedeutung eines Themas ist (Bassili, 1993; Visser, 1998). Und je höher die Bedeutung des Themas ist, umso lieber beschäftigt man sich mit diesem, was auch die Teilnahme an Befragungen einschliesst (Bizer et al., 2004).

Eine andere Erklärung im Sinne der Rational-Choice-Theorie könnte sein, dass je eher Wissen verfügbar ist, desto einfacher ist es zu antworten, was zu einem höheren Response führen kann. Dass die leicht Verfügbarkeit von Informationen mindestens das Antwortverhalten beeinflusst, ist als Phänomen des *availability bias* bekannt (Schwarz und Vaughn, 2002). Davon abgesehen ist es sehr plausibel anzunehmen, dass eine Person nicht antwortet, wenn die abgefragten Informationen erst aufwändig beschafft werden müssen. Ein Beispiel ist der hohe Nonrespone bei Fragen zur Fläche der Wohnung in der Schweiz, die den Bewohnern oft unbekannt ist und daher erst nachgeschlagen oder gar vermessen werden müsste. Es bleibt zu vermuten, dass es das Phänomen auch gibt, wenn die Befragten lediglich durch das Thema der Befragung potentielle Schwierigkeiten bei der Informationsbeschaffung antizipieren.

Der Wissenstand könnte zwar relativ leicht mit einer Selbsteinschätzungsfrage erhoben werden (Davidson et al., 1985; Cacioppo und Petty, 1980), im Rahmen der experimentellen Befragung war dies allerdings nicht möglich, da es kein eigentliches orginäres Befragungsthema gab. Nach Fragen zur Persönlichkeit und zu Werten kann nicht gefragt werden, wie gut sich die Befragten mit den Themen der Befragung auskennen.

Mittelschichtbias Ein häufig erwähnter Bias ist der Mittelschichtbias. Hartmann und Schimpel-Neimanns (1992a,b, 1993) können zeigen, dass es sich beim Mittelschichtbias um einen Bildungsbias dergestalt handelt, dass diejenigen mit hohem Bildungsabschluss eine höhere Wahrscheinlichkeit haben, an einer Befragung teilzunehmen. Und diese Gruppe der Gutausgebildeten ist auch besonders häufig in der

Mittelschicht vertreten. Beim Beispiel «Stellung im Beruf» als interessierender variable, kann kein Mittelschichtbias mehr nachgewiesen werden, wenn man die Verzerrung durch Bildung berücksichtigt.

Die Existenz eines Mittelschichtbias ist also nicht unumstritten. Insbesondere das Argument, dass Personen, die eine relative hohe Arbeitsbelastung haben, weniger schnell bereit sind, an Befragungen teilzunehmen, stimmt mindestens nicht immer. Kalfs und Saris (1998), Robinson (1999) und Bonke und Fallesen (2010) zeigen, dass gerade Personen mit hohem Arbeitsaufwand leichter zu motivieren sind. Ihre Beobachtungen stehen aber im Gegensatz zu denen der oben genannten Autoren und auch beispielsweise zu Groves und Couper (1998).

Erfahrungen mit Interviews Schlechte Erfahrungen mit Befragungen können zu gösserem Misstrauen und damit zu einer sinkenden Teilnahmebereitschaft führen (Groves et al., 1992; DeMaio, 1980). Das Markt- und Meinungsforschungsinstitut *forsa* führt jährlich eine Befragung zur Akzeptanz von Befragungen in Deutschland durch. 2006 geben fast zwei Drittel der Befragten an, bereits von einer Telemarketingagentur angerufen worden zu sein. 40% dieser Anrufe waren als Interviews getarnt (Forsa, 2006; Schnauber und Daschmann, 2008)[25]. Sheets et al. (1974) zeigen in einem Experiment, dass solche negativen Erfahrungen zu sinkenden Teilnahmebereitschaft bei folgenden Kontaktaufnahmen führen können.

Aber auch abgesehen von negativen Erfahrungen kann gezeigt werden, dass die zunehmende Häufigkeit von Anfragen zu Befragungen die Teilnahmebereitschaft senkt (Nederhof (1987),Goyder (1987),Groves et al. (1992),
Groves und Couper (1998)). In der experimentellen Befragung soll darauf verzichtet werden, nach der Häufigkeit von Einladungen zu Befragungen zu fragen. Da es sich bei den Befragten um Panelisten

25 Leider konnten keine vergleichbaren Zahlen für den Onlinebereich gefunden werden. Sehr häufig werden im Internet (und Softwarebereich) persönliche Daten der Nutzer erhoben, um diese dann kommerziell zu nutzen. Wie oft solche zu Befragungen vergleichbaren Angaben zu negativen Erfahrungen bei den Nutzern führen, müsste durch eine Befragung geklärt werden. Existierende Befragungen konnten nicht gefunden werden.

handelt, wird es für jede befragte Person eine «Panelgeschichte» geben, d. h. es ist bekannt, wie häufig eine Person mindestens durch den Panelbetreiber angefragt wurde. Ausserdem wird in einem sogenannten Basisinterview jede Person gefragt, ob sie auch an anderen Panels teilnimmt. Die Häufigkeit der Teilnahme ist damit langfristig aus technischen Gründen bekannt und muss nicht erfragt werden.

Zusätzlich ist es so, dass dieser Zusammenhang nicht unumstritten ist. So finden Sharp und Frankel (1983) oder auch Stocké und Langfeldt (2003a) keinen Zusammenhang zwischen der Häufigkeit des Anfrage und der Teilnahmebereitschaft.

Sonstiges Ein weiterer Faktor, der die Teilnahmewahrscheinlichkeit beeinflussen kann, ist die wahrgenommene Anonymität (Esser, 1973, 1986). Aus Platzgründen konnte keine Frage zu diesem Aspekt mehr aufgenommen werden.

Die Teilnahme scheitert gelegentlich an der Sorge vor einem Eindringen in die Privatsphäre (DeMaio, 1980; Smith, 1984). Es kann vermutet werden, dass Personen, die bereit sind, an einem Panel teilzunehmen, diese Sorge nicht haben. Zur Gewichtung bzw. zum Matching mittels abgeleiteter *propensity scores* ist die Frage nach der Anonymität daher (wieder vermutlich) nicht geeignet.

5 Die experimentelle Befragung

► Dieses eher kurze Kapitel beschreibt die experimentelle Befragung. Es führt kurz in das der Arbeit zu Grunde liegende Projekt ein und erläutert das Vorgehen bei der experimentellen Befragung. Die experimentelle Befragung wurde zusammen mit dem Marktforschungsunternehmen LINK im Rahmen des KTI geförderten Projekts «Online Panel Qualität» durchgeführt. ◄

5.1 Das Projekt Online Panel Qualität

LINK Web-Panel

LINK unterhält ein Online-Panel, mit dessen Hilfe insbesondere Marktforschungsstudien durchgeführt werden. LINK rekrutiert in regelmässigen Abständen neue Mitglieder für das Panel. Die experimentelle Befragung wurde im Rahmen einer solchen Rekrutierungswelle durchgeführt.

In einem Random-Quota Verfahren werden Befragte mittels Computer Assisted Telefone Interview (CATI) gefragt, ob sie bereit sind, an einem Panel teilzunehmen. Random-Quota bedeutet, dass die Auswahl der Haushalte zufällig per Random-Digit-Dialing vorgenommen wurde, die Auswahl der Person im Haushalt aber nach vorgegebenen Quoten, entsprechend bekannter Verteilungen. Die Auswahlzellen bestehen aus fünf Altersgruppen und dem Geschlecht.

Die Grundgesamtheit ist die Schweizer Wohnbevölkerung der deutsch- und französischsprachigen Schweiz.

Definition der zu ziehenden Stichprobe

Es wurde festgelegt, dass mindestens 400 Personen erreicht werden sollen, die zusagen, am Panel zu partizipieren. Diese müssen nicht aktiv sein. Aktiv ist ein Panelist, wenn er oder sie nicht nur zugesagt hat, am Panel zu partizipieren, sondern tatsächlich auch mindestens die erste, so genannte Basisbefragung, ausgefüllt hat. Desweiteren sollten mindestens 200 Personen erreicht werden, die zwar zugesagt haben, aber dann nicht auf die erste Einladung zur Befragung reagiert haben.

Wiederum 200 Personen sollten befragt werden, die zwar als Kandidaten in Frage gekommen sind, aber von vorne herein eine Teilnahme abgelehnt haben. Als Kandidaten gelten bei LINK alle, die mindestens einmal in der Woche das Internet benutzen und nicht zu «sensiblen» Berufsgruppen gehören, d. h. in der Marktforschung, Werbung oder als Journalisten tätig sind. Diese Klausel entspricht den LINK Gepflogenheiten. Und viertens sollten nochmals 200 Personen erreicht werden, die nicht als Kandidaten in Frage gekommen sind.

Realisierte Stichprobe

Tatsächlich wurde eine Stichprobe von 2'543 Befragten realisiert. Um die vorgegebenen Mindestzahlen zu erreichen, wurde ein so genannter «Boost» gestartet. Dabei werden aus Kostengründen zunächst die für den Ausschluss relevanten Fragen gestellt und nur bei positiver Antwort werden die restlichen Fragen gestellt. 475 Befragte müssen daher von allen Befragten ausgeschlossen werden, da sie im Boostteil der Befragung entweder das Internet nicht mindestens einmal in der Woche benutzen oder in der Marktforschung tätig sind. Die Grösse der Stichprobe beträgt daher netto 2'068. Die folgende Abbildung 5.1 soll dies nochmals verdeutlichen. In Klammern ist die jeweilige Fallzahl aufgeführt.

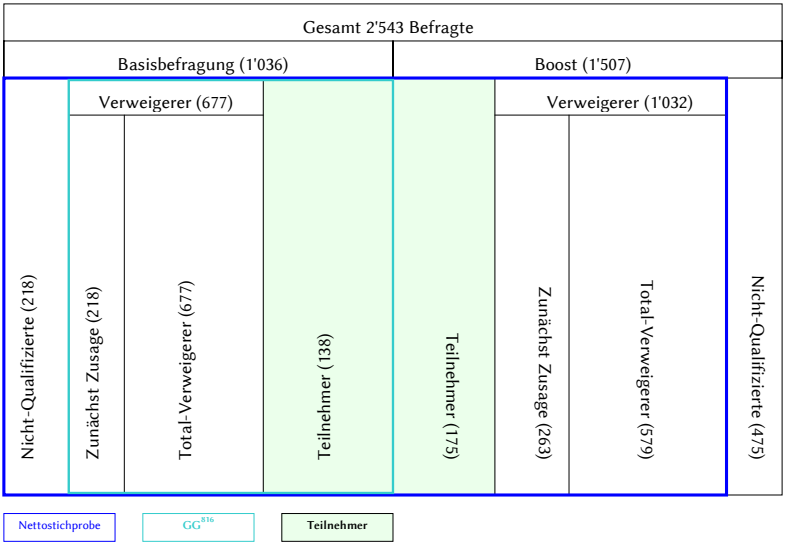


Abbildung 5.1: Umfang der Stichprobe

Der vollständige Fragebogen auf Deutsch und Französisch findet sich im Anhang, Abschnitt C, S. 241ff. Der Aufbau ist so, dass zunächst die meisten der im letzten Abschnitt beschriebenen Fragen gestellt wurden[1]. Dann folgt die Frage nach der Eignung (Internetnutzung) und dann, falls diese vorliegt, nach der Bereitschaft, am Panel teilzunehmen. Die Variablen, die helfen sollen zu erklären, warum jemand an einem Panel teilnimmt, wird also allen, die sich an der CATI Befragung beteiligen, gestellt.

Zeitraum

Die Befragung wurde am 25. Juni 2009 gestartet. Die deutschsprachigen Befragungen wurden vom LINK Telefonlabor in Zürich aus durchgeführt, die französischsprachigen im LINK Telefonlabor in Lausanne.

Da die Befragung in das regelmässige Rekrutierungsinterview von LINK implementiert wurde, mussten einige stilistische Kompromisse vorgenom-

1 Lediglich die Frage nach der Höhe des Einkommens wird ganz am Ende der Befragung gestellt. Erfahrungsgemäss ist die Verweigerungsrate bei dieser Frage besonders hoch. Um die Befragten nicht in eine «unkooperative Stimmung» zu versetzen, wurde diese Frage erst nach der Frage zur Bereitschaft gestellt.

men werden. LINK führt Telefon-Interviews prinzipiell in Schweizer Mundart durch. Das erhöht das Verständins der Befragten und ihre Teilnahmebereitschaft. Ausserdem fällt es den Interviewern leichter, sich in Mundart zu artikulieren.

Praktische Umsetzung

Allerdings entsteht ein Problem dadurch, dass durch unterschiedliche Interviewereffekte die Varianz der Schätzer erhöht wird (Lipps, 2007; van der Zouwen et al., 2009). So können z. B. letztere zeigen, dass es im Vergleich zwischen zwei Gruppen von Befragten zu unterschiedlichem Antwortverhalten kommt, wenn die Interviewer die Fragen einmal frei formulieren dürfen und in der Vergleichsgruppe die Fragen genau vorlesen müssen.

Da Schweizerdeutsch keine Schriftsprache ist, wurde die Fragen in Hochdeutsch formuliert und verschriftlicht. Nur kleine mundartliche Einsprängsel wurden gemäss LINK Usus verwendet, wie z. B. «ä chli», «öppis». Die Fragen selbst wurden von den Interviewern mundartlich gestellt, d. h. die Fragen wurden je nach Interviewer und seiner regionalen Herkunft etwas anders gestellt. In der persönlichen Überwachung zeigt sich zwar, dass die Unterschiede in der Formulierung der Fragen nicht sehr gross waren – es gab eine ausführliche Schulung und ein Training – aber nichtsdestoweniger muss mit Interviewereffekten gerechnet werden. Noch deutlicher werden sich die Interviewereffekte dadurch durchschlagen, dass die Interviewer angehalten waren, die Zahl der «weiss nicht» Antworten soweit wie möglich zu reduzieren. Dadurch kommt es immer wieder vor, dass die Interviewer eine Frage durch ein selbst gewähltes Beispiel illustrieren [2]. Die Reliabilität der Interviews ist damit etwas eingeschränkt, was aber ein generelles Problem bei CATI Befragungen ist.

5.2 Idee der Befragung

Das Ziel der Befragung ist es zu testen, ob eine Biasreduktion durch Anwendung von Gewichten, die aus *propensity scores* abgeleitet wurden, möglich

2 Es war während der Feldzeit immer möglich, bei LINK den Interviewern unbeobachtet zuzuhören. Es ist gelegentlich vorgekommen, dass die Befragten Schwierigkeiten bei den Selbsteinschätzungsfragen mittels Schwartz- und Big Five Skalen hatten. In solchen Fällen haben die Interviewer häufig Beispielsituationen konstruiert, um die Beantwortung zu erleichtern. Z. B. wurde die Frage «Ich bin jemand, dem im Leben Abwechslung wichtig ist» mit «Kaufen Sie z B. immer die gleiche Brotsorte» illustriert. Sicherlich ist dann die Antwort auf die Frage jeweils sehr abhängig vom gewählten Beispiel.

ist und falls dem so ist, zu testen, welche Variablen dafür am geeignetsten sind. Abbildung 5.2 zeigt, an welchen Stellen in der Auswahl der Befragten es zu einem Bias kommen kann und wie dieser durch die Entwicklung geeigneter Gewichte reduziert werden soll.

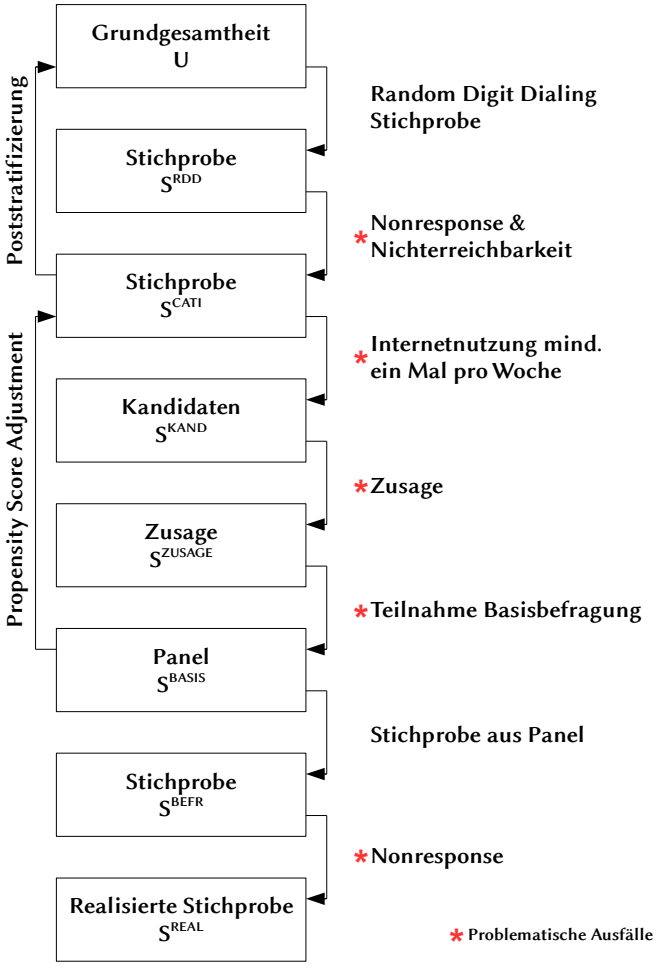


Abbildung 5.2: Ausfallprozesse und Gewichtungen

Ein wichtiger Referenzpunkt bei der Befragung ist die realisierte CATI-Stichprobe. Eine Abschätzung inwieweit die Teilnahme an der telefonischen Befragung zu einem Bias geführt hat, ist nur durch einen Vergleich mit

externen demografischen Variablen möglich. Dazu werden die demografischen Angaben der Befragten zu Alter, Geschlecht, Erwerbsgrad, Region und Einkommen mit den Zahlen verglichen, die das Bundesamt für Statistik veröffentlicht und daraus geeignete Gewichte abgeleitet, es wird poststratifiziert.

Zeigt sich später, dass das *propensity score weighting* erfolgreicher ist, als Poststratifizierung gemäss demographischer Variablen, ist dies ein guter Indikator dafür, dass auch hier die Poststratifizierung nicht das bestmögliche Gewichtungsverfahren gewesen ist. Nichtsdestotrotz hilft die Gewichtung gemäss Poststratifizierung (vermutlich) den entstandenen Bias zu reduzieren.

Diese gewichtete Stichprobe bildet die Grundgesamtheit aus Sicht des *propensity score weightings*. Zwischen der realisierten Online-Stichprobe und der realisierten RDD Stichprobe liegen einige biaserzeugende Ausfallschritte: eine telefonisch befragte Person muss zunächst ein möglicher Kandidat sein, sie muss ihre Zusage geben, am Panel teilzunehmen und diese Bereitschaft dann durch ein Basisinterview unter Beweis stellen. Tatsächlich ist die Liste der Ausfallmechanismen damit noch nicht vollständig, wenngleich in der Befragung hier an dieser Stelle abgebrochen wird. Selbst wenn jemand an der Basisbefragung teilgenommen hat, bedeutet dies noch nicht, dass die Person auch nach einer gewissen Zeit an einer Befragung zu einem spezifischen Thema wieder teilnehmen wird.

Der wichtigste Grund warum dieser Schritt nicht berücksichtigt wird, ist ein zeitlicher. Die Arbeit muss termingerecht abgeschlossen werden. Ein Argument dafür, dass dies unproblematisch ist, ist dass man die Basisbefragung als eigenständige Befragung interpretieren kann und sie damit schon das Ende der Kette der Ausfallmechanismen bildet. Die aus den *propensity scores* abgeleitete Gewichtung ist auf der Ebene der Basisbefragung daher sicherlich vernünftig. Das gilt aber nicht für spätere Befragungen, da nicht alle «Panelanfänger» die gleiche Wahrscheinlichkeit haben, aus dem Panel herauszufallen. Beim vorliegenden Design wird daher die spätere Panelmortalität nicht berücksichtigt.

Einschränkung

Problematisch ist dieses Vorgehen insofern, als dass möglicherweise nicht die richtigen Variablen identifiziert werden, um eine Panelmortalität gut vorhersagen zu können und die damit ein wichtiger Teil der Bestimmung der *propensity scores* sein müssten.

Intuitiv besteht allerdings die Hoffnung, dass die Neigung aus dem Panel

auszutreten, psychologisch und demographisch sehr eng verwandt ist mit der Bereitschaft, dem Panel beizutreten und die erste Befragung zu beantworten. Sollte dies der Fall sein, wäre die Auswahl der Kovariate schon mit deren Bestimmung auf der Ebene der Basisbefragung abgeschlossen. Näheres können aber erst zukünftige Untersuchungen in ein paar Jahren zeigen.

Kontrollvariablen

Um zu überprüfen, ob die abgeleiteten Gewichte tatsächlich zu einer Biasredunktion geführt haben, wurden Kontrollfragen gestellt. Mit ihrer Hilfe soll gezeigt werden, ob tatsächlich eine Verbesserung der Qualität der erhaltenen Informationen vorgenommen werden konnte.

K1 Besitzen Sie persönlich eines oder mehrere von den folgenden Abonnementen für den öffentlichen Verkehr?

- Halbtax-Abo
- Generalabonnement (GA)
- Verbund-Abonnement für Ihre Region
- Strecken-Abonnement (Monats- oder Jahres-Streckenabonnement)
- Mehrfahrtenkarte
- Gleis 7
- Anderes Abonnement
- Ich besitze keines dieser Abonnements.

K2 Welches ist Ihr Haupteinkaufsort für Lebensmittel und Artikel vom täglichen Bedarf? (Ohne Nennung zu kodieren als Coop, Migros, Denner und andere.)

Die Verteilung der Kontrollvariable kann bei den CATI-Teilnehmern verglichen werden mit der Verteilung bei den Web-Panelisten vor und nach der Anwendung des PSA. Es kann dann abgelesen werden, ob eine Biasreduktion mindestens bei den Kontrollvariablen erreicht werden konnte. Ausserdem erlaubt das Vorgehen den Vergleich verschiedener alternativer Gewichtungskonzepte, wie z. B. dem Vergleich zwischen verschiedenen PSA-Modellen und dem Vergleich zwischen PSA und Poststratifizierung.

6 Vergleich der Web- und CATI-Befragung

► Ein Aspekt dieser Arbeit ist es neben der Biasreduktion auch, das Teilnahmeverhalten besser zu verstehen. In einem ersten Auswertungsschritt, wurden die Teilnehmer und Verweigerer am Web-Panel bezüglich der erhobenen Variablen verglichen. Ein solcher Vergleich sagt natürlich nur etwas über einen Ausschnitt aus dem Ausfallprozess aus. Wie schon früher (Kapitel 2, insbesondere auch Abbildung 2.1 auf S. 22) gezeigt, ist der Ausfallprozess zwischen Rekrutierung und Web-Befragung lang. Ein Schritt der Rekrutierung ist Motivation der Befragten, am Panel teilzunehmen. Dies wird via CATI-Befragung vorgenommen. Der Vergleicht erlaubt also eine Analyse des Ausfallprozesses zwischen CATI- und Web-Befragung. Die in diesem Abschnitt zu beantwortende Frage lautet also: Führt der Wechsel des Befragungsmediums zu einem Bias, der mit den hier erhobenen Kovariaten messbar ist? Um den Vergleich zu vereinfachen, wurde er gemäss den im Kapitel 4 vorgeschlagenen Theorien gegliedert. ◄

Zunächst folgt eine Tabelle der erhobenen Variablen. Sie soll für spätere Abschnitte und Kapitel als Referenz dienen, da häufig kein Raum sein wird, die Variablen mit der ganzen zu Grunde liegenden Frage zu bezeichnen.

Tabelle 6.1: Variablenliste

Demografische Variablen	
f90300	Was ist Ihr Zivilstand ?
N00105	Darf ich fragen wie alt Sie sind (Kategorisiert)
N85200	Wie oft haben Sie in den letzten 12 Monaten etwas

Fortsetzung nächste Seite . . .

Tabelle 6.1: Variablenliste

	online im Internet gekauft?
N90401	A) Kinder bis 5 Jahre
N90402	B) Kinder 6- 9 Jahre
N90403	C) Kinder 10-14 Jahre
N90404	D) Jugendliche 15-19 Jahre
N90405	E) Erwachsene 20-64 Jahre
N90406	F) Erwachsene ab 65 Jahre
anzErw	Anzahl Erwachsene (N90405 + N90406)
HHgroesse	Haushaltsgrösse (Summe N90401 bis N90406)
f91100	Welche Schule haben Sie zuletzt besucht?
f91300	Persönliches Einkommen (6er Skala)
f91600	WEMF Region
f00110	Geschlecht der befragten Person
f00140	Sind Sie voll, teilweise oder nicht erwerbstätig?
f03200	Seit wann wohnen Sie am jetzigen Wohnort (5er Skala)
Schwartz Skala	
f03401	Ich jemand dem es wichtig ist, seine Fähigkeiten zu zeigen.
f03402	. . .in einer sicheren Umgebung zu leben.
f03403	. . .immer wieder neue Aktivitäten auszuprobieren.
f03404	. . .im Leben Abwechslung zu haben.
f03405	. . .sich immer an Regeln zu halten, selbst dann, wenn es niemand sieht
f03406	. . .zurückhaltend und bescheiden zu sein.
f03407	. . .Spass zu haben.
f03408	. . .den Menschen um mir herum zu helfen.
f03409	. . .sich immer richtig zu verhalten und nicht den Eindruck zu erwecken, mich falsch zu verhalten
f03410	. . .von anderen respektiert zu werden.
Big Five Skala	
f03301	Ich bin eine Person, wo gründlich schafft.

Fortsetzung nächste Seite . . .

Tabelle 6.1: Variablenliste

f03302	...wo kommunikativ und gesprächig ist
f03303	...wo manchmal ä chli grob zu anderen ist
f03304	...wo originell ist und Ideen einbringt
f03305	...wo sich oft Sorgen macht
f03306	...wo verzeihen kann
f03307	...wo eher faul ist
f03308	...wo aus sich herausgehen kann und gesellig ist
f03309	...wo künstlerische Erfahrungen schätzt
f03310	...wo leicht nervös wird
f03311	...wo Aufgaben wirksam und effizient erledigt
f03312	...wo zurückhaltend ist
f03313	...wo rücksichtsvoll und freundlich mit anderen umgeht
f03314	...wo eine lebhafte Phantasie hat
f03315	...wo entspannt ist und mit Stress gut umgehen kann
Rational Choice Fragen	
f03501	Marktforschungsunternehmen behandeln die Daten vertraulich.
f03502	Umfragen bringen Abwechslung und sind interessant.
f03503	Bei Umfragen wird häufig öppes gefragt, wo niemand öppes angeht.
f03504	Ich bin bereit, über meine persönlichen Gewohnheiten auch mit anderen zu sprechen.
f03505	Marktforschung ist für die Gesellschaft wichtig und sinnvoll.
f85100	Wie häufig etwa nutzen Sie das Internet für private Zwecke?
f85300	Haben Sie zu Hause einen Breitband-Internet-Anschluss mit ADSL oder Kabelmodem oder nicht?
f86110	Das es für die Teilnahme eine Belohnung gibt, ist für mich ein wichtiger Anreiz
Testvariablen	

Fortsetzung nächste Seite ...

Tabelle 6.1: Variablenliste

f03001	Haupteinkaufsort für Lebensmittel ist Migros
f03002	. . .ist Coop
f03003	. . .ist Denner
f03004	. . .ist anderer Supermarkt
f03101	Ich besitze ein Halbtax
f03102	. . .ein Generalabonnement
f03102	. . .ein Verbund Abonnement
f03102	. . .ein Strecken Abonnement
f03102	. . .eine Mehrfahrtenkarte
f03102	. . .ein Gleis 7 Abonnement
f03102	. . .ein anderes Abonnment
f03102	. . .kein genanntes Abonnement
Paradaten	
noNA	Anzahl fehlender Werte bei allen Variablen.

Es ist natürlich per se interessant, sich die Unterschiede zwischen den freiwilligen Teilnehmern an einem Web-Panel mit denjenigen zu vergleichen, die zwar an einer CATI-Befragung teilnehmen, nicht jedoch am Web-Panel. Der Vergleich ist auch eine hilfreiche Vorbereitung für die späteren Schritte, bei denen die Teilnahmewahrscheinlichkeit geschätzt werden soll. Da die Schätzung multivariat erfolgen wird, kann es sich allerdings nur um Hinweise auf hilfreiche Variablen handeln: Die Existenz oder das Fehlen eines Unterschieds ist in Bezug auf die multivariate Schätzung noch kein Bewies dafür, dass die Variable eine hohe oder niedrige Prognosekraft besitzt. Die folgenden Ausführungen beziehen sich immer auf alle Befragten. Das wird bei späteren Auswertungen nicht immer der Falls sein, aber dort gesondert vermerkt werden.

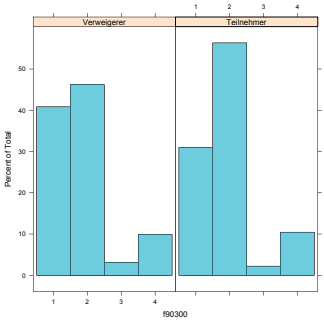
6.1 Demografische Unterschiede

Die Unterschiede in der demografischen Struktur der CATI-Befragung im Vergleich zur Population können nicht sehr stark sein, da bei der Rekrutierung ein sogenanntes Random-Quota Verfahren gewählt wurde. Das bedeutet,

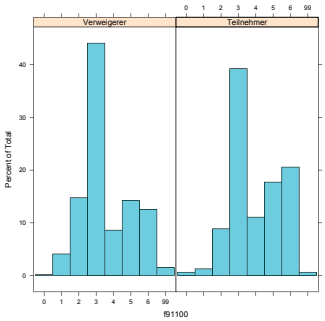
dass die Haushalte zwar zufällig ausgewählt wurden, allerdings eine «Quotenverfahren» bei der Auswahl der einzelnen Person aus dem Haushalt angewendet wurde. (Existiert im gezogenen Haushalt keine Person mit dem geforderten Quotenmerkmal, wird der Haushalt verworfen.) Die zugrundeliegenden Variablen sind insbesondere das Alter und Geschlecht. Es ist bei diesen Variablen daher mit mit einer kleinen Abweichung zu rechnen. Da die Parameter der Verteilungen der Quotierungsvariablen allerdings für die Grundgesamtheit bekannt sind, handelt es sich eher um eine geschichtete Stichprobe.

Nichtsdestoweniger ist der Vergleich der demografischen Variablen zwischen den Teilnehmern am Panel und der Teilnehmern an der Online-Befragung interessant.

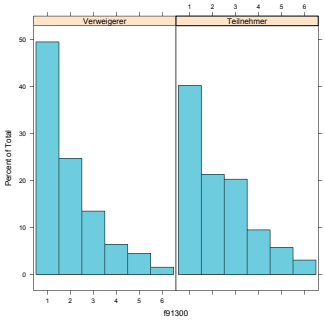
Die folgenden Abbidungen sind so zu lesen, dass immer die Verteilungen der jeweiligen Variable für die Teilnehmer am Panel den restlichen CATI-Befragten gegenübergestellt werden. Letztere werden *Verweigerer* genannt, obwohl auch diejenigen zu dieser Gruppe gehören, die z. B. wegen zu geringer Internetnutzung nicht in Frage kommen. Damit die einzelnen Abbildungen nicht überladen werden, sind sie nur mit Ziffern für die Ausprägungen beschriftet. Jeweils rechts neben der Abbildung sind daher nochmals neben dem Variablennamen auch die Ausprägungen, immer beginnend bei 1, aufgelistet. 99 steht jeweils für «keine Antwort».



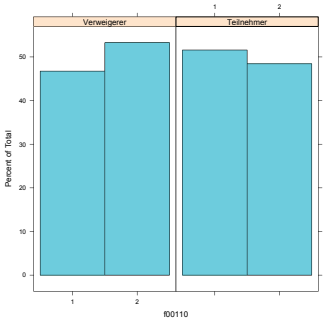
`f90300` (Zivilstand)
ledig, verheiratet, verwitwet, geschieden, k. A.



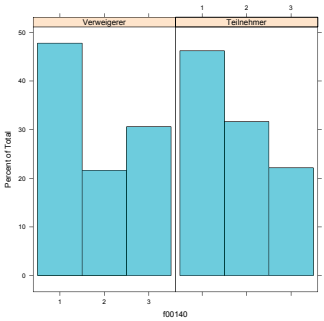
`f91100` (Höchste Ausbildung)
keine, Primars., Sekundars. Berufss., Mittels., Technikum/FH, Uni



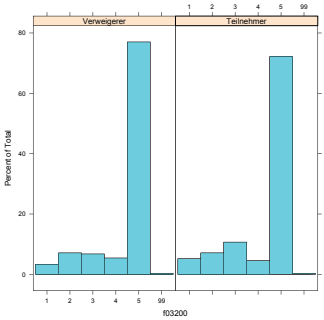
`f91300` (Bruttoeinkommen monatlich)
in SFr: <4'000, 4'001–6'000, 6'001–8'000, 8'001–10'000, 10'000–15'000, >15'000



`f00110` (Geschlecht)
Mann, Frau



`f00140` (Erwerbstätigkeit)
voll, teilweise, nicht erwerbstätig



`f03200` (Wohndauer)
seit diesem Jahr, letztem Jahr, vorletztem Jahr, seit ca. 3 Jahren, länger

Panelisten sind also eher verheiratet, Verweigerer haben einen höheren Ledigenanteil. Panelisten sind etwas besser gebildet und haben ein höheres Einkommen. Der Männeranteil ist bei ihnen etwas höher. Alle drei Variablen korrelieren jeweils miteinander.

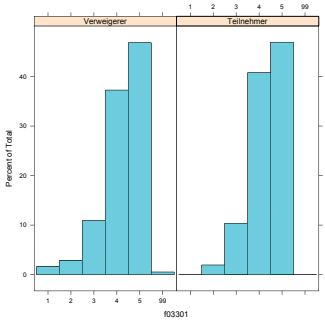
Der Anteil der voll Erwerbstätigen ist in etwa gleich hoch, allerdings ist der Anteil der Teilzeitbeschäftigten bei den Panelisten höher und der Anteil der nicht-Beschäftigten entsprechend niedriger. Bei der Wohndauer in der jetzigen Wohnung – einer von LINK standardmässig erhobenen Variable – unterscheiden sich beide Gruppen nur gering.

6.2 Unterschiede in den Persönlichkeitsvariablen.

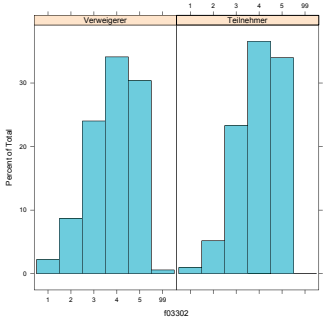
Wie in Abschnitt 4.3 (S. 70 ff.) erläutert, könnten es Unterschiede in der Persönlichkeit sein (definiert als das, was das Big Five Inventory misst), die das Teilnahmeverhalten an Web-Befragungen bestimmen. Die folgenden

Abbildungen vergleichen dem Schema des letzten Abschnitts folgend die entsprechenden Variablen zwischen Teilnehmern und Verweigerern.

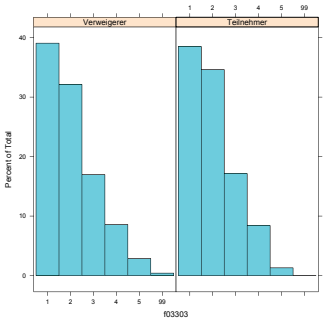
Die Skala ist bei allen Fragen: 1 = *trifft auf mich überhaupt nicht zu* bis 5 = *trifft auf mich voll und ganz zu*. Es handelt sich um eine endpunktbeschriftete Skala, die Auspägungen 2 bis 4 sind also nicht explizit beschriftet. Alle Fragen sind Selbsteinschätzungen der Form, *Ich bin eine Person, wo...*



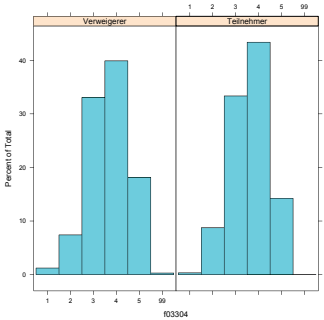
`f03301` (...gründlich schafft.)



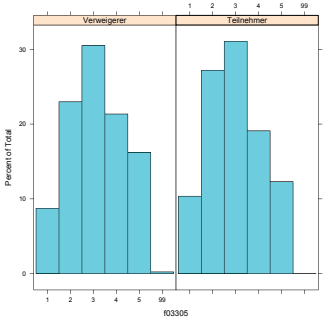
`f03302` (...kommunikativ und gesprächig ist)



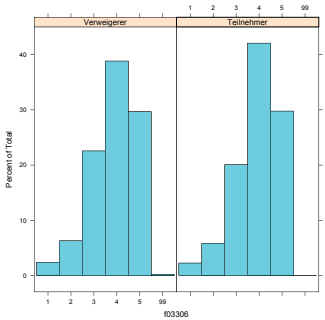
`f03303` (. . .manchmal ä chli grob zu anderen ist)



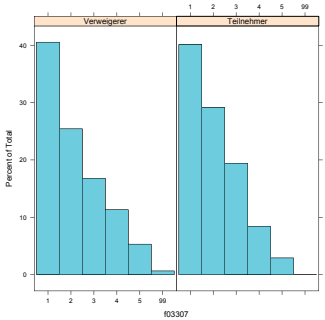
`f03304` (. . .originell ist und Ideen einbringt)



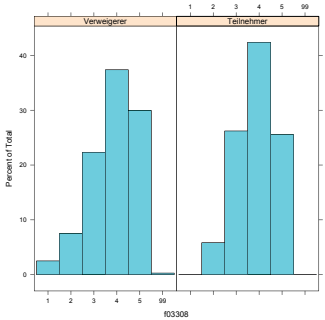
`f03305` (. . .sich oft Sorgen macht)



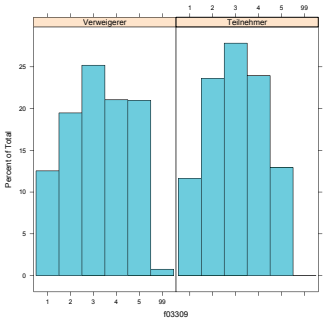
`f03306` (. . .verzeihen kann)



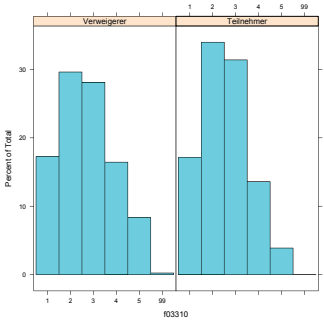
`f03307` (. . .eher faul ist)



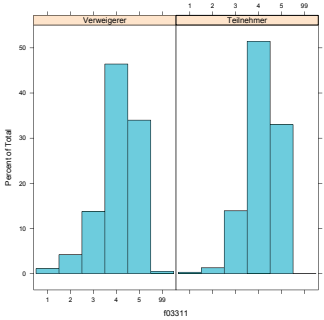
`f03308` (. . .aus sich herausgehen kann und gesellig ist)



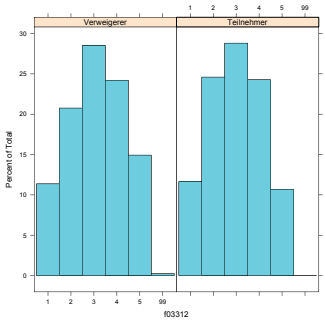
`f03309` (. . .künstlerische Erfahrungen schätzt)



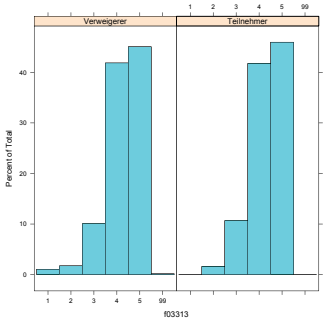
`f03310` (. . .leicht nervös wird)



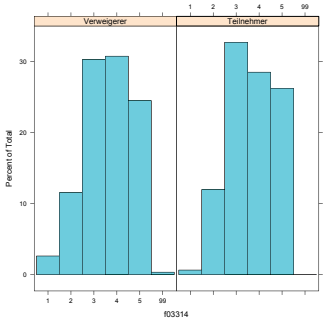
`f03311` (. . .Aufgaben wirksam und effizient erledigt)



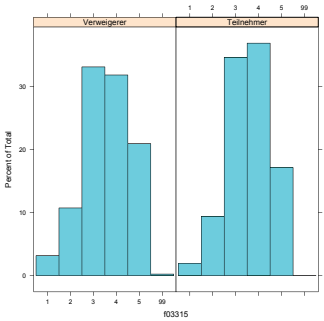
f03312 (. . .zurückhaltend ist)



f03313 (. . .rücksichtsvoll und freundlich mit anderen umgeht)



f03314 (. . .eine lebhafte Phantasie hat)



`f03315` (. . .entspannt ist und mit Stress gut umgehen kann)

Es gibt bei keiner einzigen Persönlichkeitsvariable aus dem Big Five Inventory einen signifikanten Unterschied zwischen CATI-Teilnehmenden und Web-Panelisten. Ignoriert man für einen Augenblick den Umstand, dass es multidimensional doch Unterschiede zwischen beiden Gruppen geben kann (die es nicht gibt, siehe Abschnitt 8), ist dies ein interessantes Ergebnis.

Offensichtlich sind es keine Persönlichkeitsvariablen, die die Bereitschaft beeinflussen an einer Web-Befragung teilzunehmen. Jedenfalls gilt dies für alle, die bereits bereit waren, an einer CATI-Befragung teilzunehmen. Es ist möglich, dass es Persönlichkeitsmerkmale sind, die die Teilnahme an Befragungen im allgemeinen steuern. Das lässt sich aber mit dem hier gewählten Untersuchungsdesign nicht untersuchen.

Zunächst existieren keine Informationen über die Verteilung der Persönlichkeitsvariablen in der Grundgesamtheit. Auch ohne diese könnte man argumentieren, dass wenn die entsprechenden Eigenschaften das Verhalten nicht vollständig determinieren, es einige in der CATI-Befragung geben sollte, die ein Persönlichkeitsprofil haben, dass eher gegen eine Teilnahme an Befragungen spricht. Diese Gruppe sollte dann in der Web-Befragung in sehr viel kleinerer Zahl vertreten sein. Ein solcher Unterschied lässt sich aber nicht feststellen, was als (schwaches) Indiz dafür gewertet wird, dass die Teilnahme an Befragungen nicht durch die (erhobenen) Persönlichkeitsvariablen bestimmt wird.

«Schwach» ist das Indiz aus zwei Gründen:

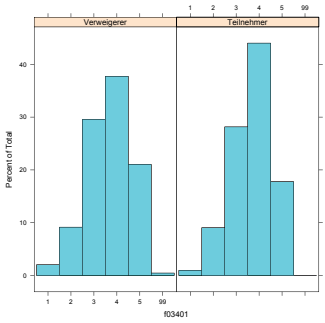
1. Es ist möglich, dass die Stichprobe zu klein ist, um einen solchen Unterschied zwischen Teilnehmern und Verweigerern zu entdecken.

2. Die Persönlichkeitsvariablen könnten nur einen schwachen Einfluss auf das Teilnahmeverhalten haben.

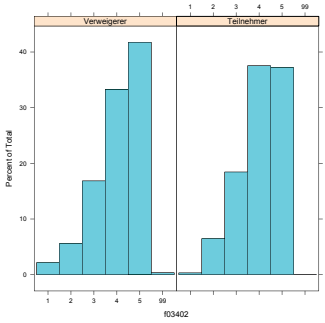
6.3 Unterschiede in den Werten

Die Werte wurden mittels Schwartz-Werte-Skala erfasst, siehe Abschnitt 4.4. Auch hier sollen die erhobenen Variablen zwischen Teilnehmern und Verweigerern verglichen werden.

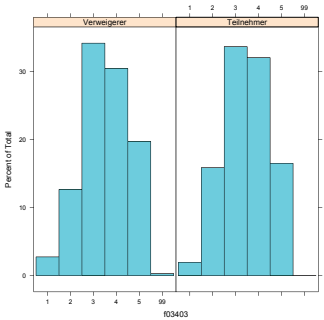
Alle Fragen haben wieder eine ähnliche Antwortskala wie die Fragen zu den Persönlichkeitsvariablen des vorangegangenen Abschnitts: 1 = *trifft auf mich überhaupt nicht zu* bis 5 = *trifft auf mich voll und ganz zu*. Die vorgegebenen Antworten beginnen jeweils mit *Ich jemand dem es wichtig ist, ...*.



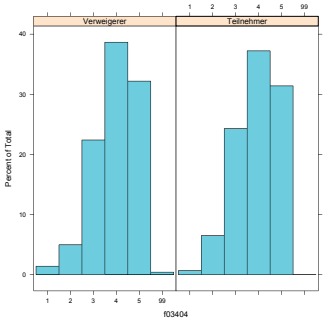
`f03401` (...seine Fähigkeiten zu zeigen.)



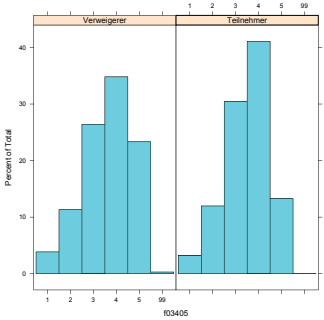
`f03402` (...in einer sicheren Umgebung zu leben.)



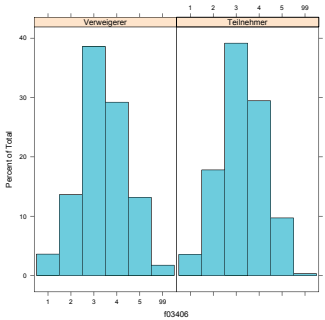
`f03403` (. . .immer wieder neue Aktivitäten auszuprobieren.)



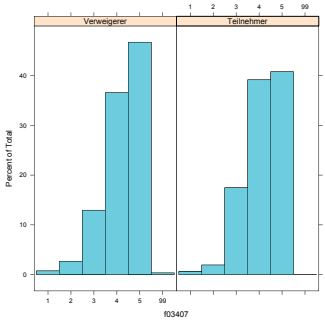
`f03404` (. . .im Leben Abwechslung zu haben.)



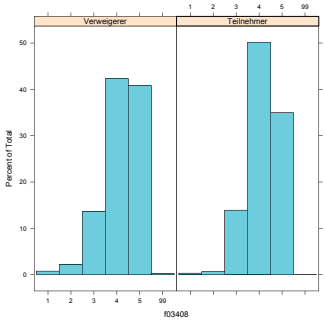
`f03405` (. . .sich immer an Regeln zu halten, selbst dann, wenn es niemand sieht.)



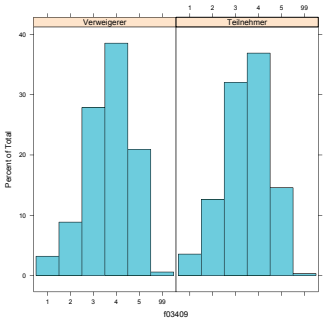
f03406 (. . .zurückhaltend und bescheiden zu sein.)



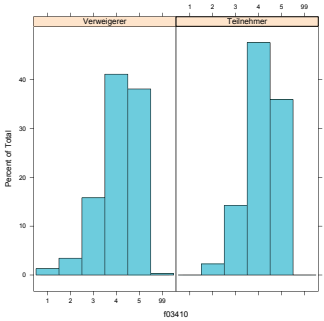
f03407 (. . .Spass zu haben.)



f03408 (. . .den Menschen um mir herum zu helfen.)



f03409 (. . .sich immer richtig zu verhalten und nicht den Eindruck zu erwecken, mich falsch zu verhalten.)



f03410 (. . .von anderen respektiert zu werden.)

Bei den Variablen, die die Werte abbilden sollen, unterscheiden sich zwei Variablen signifikant zwischen den beiden Gruppen: f03405 (. . .sich immer an Regeln zu halten, selbst dann, wenn es niemand sieht.) und f03409 (. . .sich immer richtig zu verhalten und nicht den Eindruck zu erwecken, mich falsch zu verhalten.)[1]. Beide Variablen sind inhaltlich sehr nahe verwandt.

Vermutlich nehmen Personen, die gebeten werden an einer Befragung teilzunehmen, eine Norm wahr, anderen helfen zu sollen. Diese Regel (abgefragt mit Variable f03408) scheint aber keine Rolle bei der Entscheidung zu spielen, am Panel teilzunehmen. Allerdings ist eine zweite Norm, nämlich sich an Normen und Regeln zu halten, doch wichtig bei der Entscheidung. Es genügt also offensichtlich nicht eine Norm zu antizipieren, man muss

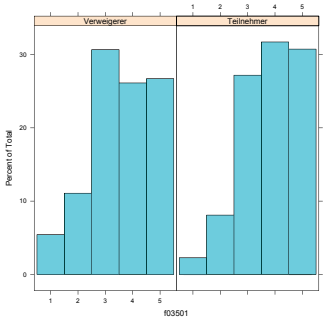
1 Getestet mit dem t-Test, Signifikanzniveau 5%.

auch der vermittelnden «Metanorm» folgen, sich an die Norm zu halten.

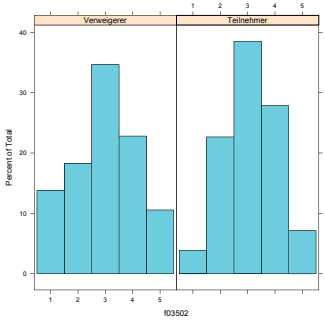
6.4 Unterschiede bei den Rational Choice Fragen

Die ersten fünf Fragen (`f03501` bis `f03505`), die sich unmittelbar als Fragen präsentieren, die aus der Theorie rationaler Wahl abgeleitet wurden, haben als Antwortskalen 1 = *Stimme überhaupt nicht zu* bis 5 = *Stimme voll und ganz zu*. Wiederum sind nur die Endpunkte beschriftet. Die gleiche Skala trifft auch auf `f86110` zu.

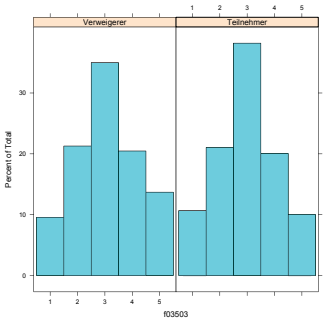
Die Fragen wurden jeweils eingeleitet von der Frage *Bitte sagen Sie uns jeweils, inwiefern Sie persönlich diesen Aussagen zustimmen.*



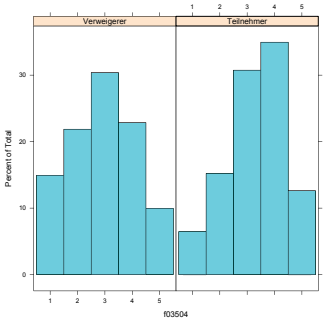
`f03501` (Marktforschungsunternehmen behandeln die Daten vertraulich.)



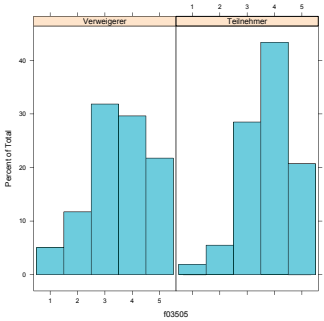
`f03502` (Umfragen bringen Abwechslung und sind interessant.)



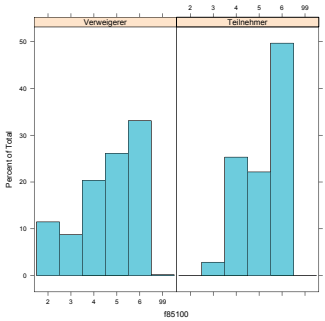
`f03503` (Bei Umfragen wird häufig öppes gefragt, wo niemand öppes angeht.)



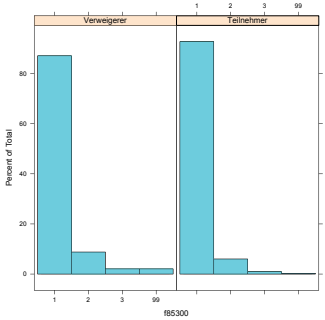
`f03504` (Ich bin bereit, über meine persönlichen Gewohnheiten auch mit anderen zu sprechen.)



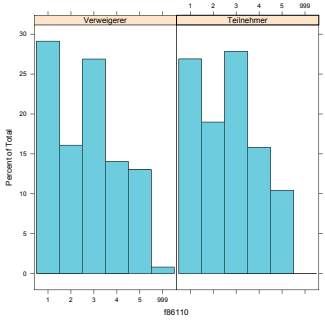
`f03505` (Marktforschung ist für die Gesellschaft wichtig und sinnvoll.)



`f85100` (Wie häufig etwa nutzen Sie das Internet für private Zwecke?)
seltener, 1× pro Woche, mehrmals pro Woche, 1× täglich, mehrmals täglich, nie



`f85300` (Haben Sie zu Hause einen Breitband-Internet-Anschluss mit ADSL oder Kabelmodem oder nicht?)
ja, keinen Breitbandanschluss, gar kein Internet



`f86110` (Das es für die Teilnahme eine Belohnung gibt, ist für mich ein wichtiger Anreiz.)

Das Marktforschungsunternehmen Daten vertraulich behandeln, glauben weder die Teilnehmer noch die Verweigerer und unterscheiden sich in ihrer Meinung auch nicht signifikant. Beide Gruppen unterscheiden sich auch nicht in ihrer Einschätzung, dass bei Befragungen häufig etwas gefragt wird,

dass niemanden etwas angeht (Vielleicht ist auch nur der Eindruck der CATI-Befragung auf beide Gruppen gleich gross gewesen und der *availability bias* ist gleich stark).

Unterschiede gibt es allerdings bei der Einschätzung, dass Umfragen abwechslungsreich und interessant sind. Die Teilnehmer halten Befragungen für abwechslungsreicher und / oder interessanter, mithin also für unterhaltender.

Beide Gruppen unterscheiden sich bei den wahrgenommenen Kosten von Befragungen in Bezug auf die Bereitschaft, über persönliche Gewohnheiten auch mit anderen zu sprechen[2].

Ausserdem halten Freiwillige Marktforschung für die Gesellschaft sinnvoller und / oder wichtiger. Bei der empfundenen Wichtigkeit der Anreize unterscheiden sich beide Gruppen nicht.

Einen offensichtlich grossen Unterschied gibt es auch bei der Häufigkeit, das Internet zu benutzen. Teilnehmer benutzen das Internet häufiger als Verweigerer. Es muss allerdings nochmals darauf hingewiesen werden, dass Link all diejenigen, die das Internet seltener als einmal in der Woche benutzen, nicht als Panelist zulässt. Es gibt also eine definitorische Verschiebung. Aber auch jenseits davon kann gezeigt werden, dass die Häufigkeit das Internet zu benutzen, eine der Variablen mit der höchsten Erklärungskraft in allen Modellen sein wird.

6.5 Fazit

Die Unterscheide zwischen den Panelisten und den Verweigerern sind insgesamt relativ gering. Nur bei wenigen Variablen lassen sich signifikante Unterschiede feststellen und selbst bei diesen Variablen ist der Unterschied nicht sehr stark. Eine einzelne Variable, die eine hohe Erklärungskraft der Teilnahme erlaubt, lässt sich nicht feststellen.

Insbesondere überraschend ist, dass die Persönlichkeitsmerkmale (eindimensional) gar keine Rolle spielen zu spielen scheinen und auch der Einfluss der Werte nur gering ist – immer vorausgesetzt, dass die Items und Skalen

2 Ob man diese Frage eher zu den Persönlichkeitsmerkmalen zählt, sie als Kosten interpretiert oder eventuell sogar zu den Fragen zählt, die die Werte betreffen, ist im Sinne dieser Arbeit unwichtig.

brauchbar sind. Die demografischen Fragen und die Fragen, die aus der Theorie rationaler Wahl abgeleitet wurden, scheinen am erfolgversprechendsten das Teilnahmeverhalten vorherzusagen zu können.

Leider unterscheiden sich auch insbesondere die Kontrollvariablen nur sehr gering zwischen den beiden Gruppen. Das ist im Sinne dieser Arbeit ungünstig, denn mit ihrer Hilfe sollte ja überprüft werden, ob die Methode der Gewichtung mittels Teilnahmewahrscheinlichkeit erfolgversprechend ist.

Der Abwechslung wegen soll der Vergleich bei den Testvariablen zum Abschluss tabellarisch (6.2) erfolgen.

Es besteht lediglich ein signifikanter Unterschied bei den Mehrfahrtenkarten und dem nicht-Besitz eines SBB-Abonnements.

Zusammenfassend lässt sich ableiten, dass da die Unterschiede zwischen beiden Befragungsmodi nur gering sind, beide Befragungsformen mindestens für die hier erhobenen Fragen äquivalent sind! Das ist überraschend, wenn man die Diskussion über die Güte von Web-Befragungen wie sie kurz in der Einleitung skizziert wurde berücksichtigt.

Die Ausfallmechanismen, zwischen der Gruppe derer, die am Telefoninter-

	A Verweigerer	nteil in Prozent Teilnehmer
Migros	63.6	64.7
Coop	55.2	44.0
Denner	8.3	7.7
Andere	21.5	18.4
Halbtax	36.7	42.4
GA	11.9	12.6
Verbund-Abo	10.5	14.2
Strecken-Abo	8.5	6.1
Mehrfahrtenkarte	10.4	17.1
Gleis 7	1.9	1.9
Andere	2.5	2.9
Keine	39.9	32.7

Tabelle 6.2: Vergleich der Testvariable zwischen Teilnehmern und Verweigerern

view teilgenommen haben und denjenigen, die aus dieser Gruppe bereit sind, auch an einem Web-Panel zu partizipieren, sind vernachlässigbar in dem Sinne, dass sie nicht zu einem Bias führen. Für alle denkbaren Variablen kann man dies selbstverständlich nicht generalisieren, aber da es sich bei den erhobenen Variablen um solche handelt, die nach allem, was bekannt ist, sehr hoch mit dem Teilnahmeverhalten korrelieren, ist die grosse Nähe beider Gruppen jedenfalls beeindruckend.

Der Vergleich zwischen den Panelisten und den CATI-Befragten kann so verstanden werden, dass es keinen Unterschied ergibt, ob Befragte CATI befragt werden oder in einem CATI-rekrutierten Web-Panel. Die Ergebnisse sprechen dagegen noch lange nicht dafür, dass jegliche Art von Web-Befragung ähnliche Ergebnisse liefert wie eine sorgfältig durchgeführte CATI-Befragung.

7 Auswahl der Kovariaten mittels Baum-Modell

► Neben der Identifizierung geeigneter Kovariaten ist die Modellierung entscheidend beim Verstehen des Teilnahmeverhaltens. In der Statistik sind viele Methoden bekannt, die eine solche Modellierung prinzipiell erlauben. Die meisten Methoden scheitern allerdings an dem Umstand, dass es bei der experimentellen Befragung sehr viele fehlende Werte gegeben hat. Auf Grund dieses Umstandes war es notwändig, möglichst ein nicht-parametrisches Verfahren zu verwenden. Hier wurden Klassifikations- und Regressionsbäume (*classification and regression trees, CART*) verwendet. Für eine Einführung in die Methode siehe Abschnitt 3.3.2, ab S. 36. Ähnlich dem Vorgehen einer Modellierung mittels Regressions-Modellen, kann auch bei dieser Methode zunächst ein Modell (also Baum) mit allen zur Verfügung stehenden Variablen bestimmt werden, um dann mittels Elimination systematisch Variablen auszuschliessen, die keine oder nur geringe Erklärungskraft für die Bestimmung der Teilnahmeneigung haben. Entsprechend diesem Vorgehen ist auch das Kapitel strukturiert. Zunächst wird in Abschnitt 7.1 ein vollständiger Baum beschrieben. Im anschliessenden Abschnitt 7.2 die Kriterien angegeben, mit deren Hilfe der dann in Abschnitt 7.3 zu beschreibende definitive Baum entwickelt wurde. ◄

Es gibt weitere nicht-parametrische Methoden, die zwar ausprobiert wurden um die Teilnahmewahrscheinlichkeit zu bestimmen, aber gescheitert sind. Zu diesen Methoden gehören insbesondere Verfahren wie EM-Algorithmen (Benaglia et al., 2009)[1] oder Kernel-Smoother (Wand und Jones,

1 Auch umgesetzt als R-Package `mixtools`, (Benaglia et al., 2009).

1994)[2]. Diese Methoden sollen im Folgenden undokumentiert bleiben.

7.1 Vollständiger Baum

Um herauszufinden, welche Variablen für die Konstruktion eines Klassifikationsbaums die grösste Bedeutung haben, wurden alle plausiblen Variablen zur Berechnung benutzt. Alle Variablen wurden als kategorial behandelt, mit Ausnahme des Alters (`N00105`) und der Anzahl fehlender Werte (`noNA`). Das ist auch deswegen geschehen, damit die fehlenden Werte als eigene Kategorie behandelt werden können.

Da die Anzahl der fehlenden Werte beim Haushaltseinkommen höher war als bei den persönlichen Einkommen, wurde nur letztere verwendet. Einige fehlende Angaben zum persönlichen Einkommen konnten mit Hilfe der Angabe zum Haushaltseinkommen ersetzt werden. Dazu wurde das Haushaltseinkommen durch die Anzahl von Personen über dem Alter von 20 Jahren[3] dividiert.

Das (Brutto-) Einkommen ist kategorisch mit einer sechsstufigen Skala[4] erhoben wurden. Um den Haushaltsdurchschnitt zu bestimmen wurde einfach der Mittelwert der jeweiligen Kategorie durch die Anzahl der Haushaltsmitglieder über 20 Jahre dividiert und gerundet. Das konnte für 64 Haushalte vorgenommen werden.

Die Berechnungen wurden mit dem R Package *rpart* (Therneau et al., 2009) durchgeführt.

Die Routine zur Bestimmung des Baums hat folgende Variablen für dessen Konstruktion des Baums beibehalten: `HHgroesse`, `N00105`, `N85200`, `f03307`, `f03308`, `f03309`, `f03311`, `f03312`, `f03315`, `f03406`, `f03407`, `f03408`, `f03409`, `f03501`, `f03502`, `f03505`, `f86110`, `f91300`, `f91600`.

169 Befragte (das entspricht 8.2 %) wurden falsch klassifiziert. Ein Auswahlkriterium für die Entwicklung eines Baumes ist die Devianz des Ursprungsknotens dividiert durch die Devianz eines fraglichen neuen Knotens.

2 Umgesetzt als R Package `KernelSmooth` (Wand und Ripley, 2010)

3 Die Altersgrenze ergibt sich aus der Art, wie LINK die Haushaltsgrösse erfasst: sie fragen nach der Anzahl Kinder bis 5 Jahren, zwischen 6 und 9 Jahren, zwischen 10 und 14 Jahren, Jugendlichen zwischen 15 und 19 Jahren, Erwachsenen zwischen 20 und 64 Jahren und Erwachsenen ab 65 Jahren.

4 In CHF: bis 4'000, zwischen 4'001 und 6'000, zwischen 6'001 und 8'000, zwischen 8'001 und 10'000, zwischen 10'001 und 15'000, mehr als 15'000.

Unterschreitet dieses Verhältnis einen vorgegebenen Wert, wird der Knoten eliminiert bzw. nicht entwickelt. Der Standardwert in rpart beträgt 0.01, ein möglicher Knoten wird also nur beibehalten, wenn dessen Devianz mindestens ein Hundertstel des Ursprungsknoten (root) beträgt. Für das vollständige Modell wurde dieser Wert auf Null gesetzt, was auch die Berücksichtigung sehr kleiner Devianzen erlaubt.

[Zum Lesen der Tabelle: ein Stern «*» bezeichnet einen Endknoten im Baum, ein so genanntes Blatt. Die Verzweigung eines Konotens wird links immer geradzahlig nummeriert und rechts ungeradzahlig.]

Tabelle 7.1: Knoten im Referenzbaum

Knoten	Split	Bedingung	Anzahl	Devianz	Teilnahme	p(kT)	p(T)	
1)	root		2068	316	kT	0.85	0.15	
2)	HHgroesse	=2,3,4,5, 6,7,9,11	1231	255	kT	0.79	0.21	
4)	f86110	=999	1	0	kT	1	0	*
5)	f86110	=1,2,3,4,5	677	255	kT	0.62	0.38	
10)	N00105	<30.5	266	53	kT	0.8	0.2	
20)	N85200	<5.5	194	28	kT	0.86	0.14	*
21)	N85200	>=5.5	72	25	kT	0.65	0.35	
42)	f03307	=2,4,5,99	35	6	kT	0.83	0.17	*
43)	f03307	=1,3	37	18	T	0.49	0.51	
86)	f03309	=1,5	12	1	kT	0.92	0.08	*
87)	f03309	=2,3,4	25	7	T	0.28	0.72	
174)	f03409	=2,5	8	2	kT	0.75	0.25	*
175)	f03409	=3,4	17	1	T	0.06	0.94	*
11)	N00105	>=30.5	411	202	kT	0.51	0.49	
22)	f03408	=1,2,3,5	222	88	kT	0.6	0.4	
44)	f03308	=1,2,5	108	31	kT	0.71	0.29	
88)	N00105	<43.5	40	4	kT	0.9	0.1	*
89)	N00105	>=43.5	68	27	kT	0.6	0.4	
178)	f91300	=2,5	17	3	kT	0.82	0.18	*
179)	f91300	=1,3,4,6	41	19	T	0.46	0.54	
358)	f03505	=5	22	7	kT	0.68	0.32	*
359)	f03505	=2,3,4	19	4	T	0.21	0.79	*
45)	f03308	=3,4	114	57	kT	0.5	0.5	
90)	f91300	=5,6	12	3	kT	0.75	0.25	*
91)	f91300	=1,2,3,4	89	37	T	0.42	0.58	
182)	f03501	=3	36	15	kT	0.58	0.42	
364)	f03315	=1,3,4,5	29	9	kT	0.69	0.31	*
365)	f03315	=2	7	1	T	0.14	0.86	*
183)	f03501	=1,2,4,5	53	16	T	0.3	0.7	*

Fortsetzung nächste Seite . . .

Tabelle 7.1: Knoten im Referenzbaum

Knoten	Split	Bedingung	Anzahl	Devianz	Teilnahme	p(kT)	p(T)	
23)	f03408	=4	185	73	T	0.39	0.61	
46)	f03502	=1,5	23	7	kT	0.7	0.3	*
47)	f03502	=2,3,4	162	57	T	0.35	0.65	
94)	f91600	=1,2,4	123	51	T	0.41	0.59	
188)	N00105	<41.5	48	20	kT	0.58	0.42	
376)	N85200	<22	39	12	kT	0.69	0.31	*
377)	N85200	>=22	9	1	T	0.11	0.89	*
189)	N00105	>=41.5	75	23	T	0.31	0.69	
378)	f91300	=2,4,5	37	17	T	0.46	0.54	
756)	f03406	=4,5	15	4	kT	0.73	0.27	*
757)	f03406	=2,3	22	6	T	0.27	0.73	*
379)	f91300	=1,3,6	36	5	T	0.14	0.86	*
95)	f91600	=3,99	39	6	T	0.15	0.85	*
3)	HHgroesse	=1,8	258	61	kT	0.76	0.24	
6)	f86110	=	0	0	kT	0	0	*
7)	f86110	=1,2,3,4,5	126	61	kT	0.52	0.48	
14)	f03407	=4,5	97	40	kT	0.59	0.41	
28)	f03311	=4	44	12	kT	0.73	0.27	*
29)	f03311	=3,5	53	25	T	0.47	0.53	
58)	f03312	=3,5	16	4	kT	0.75	0.25	*
59)	f03312	=1,2,4	37	13	T	0.35	0.65	
118)	f03409	=1,2,5	12	4	kT	0.67	0.33	*
119)	f03409	=3,4	25	5	T	0.2	0.8	*
15)	f03407	=2,3	21	5	T	0.24	0.76	*

* bezeichnet Endknoten

Bei dem knoten 6) / 7) ist es offenbar zu einem Fehler der routine gekommen, da es sich eigentlich nicht um einen Knoten handelt, da alle in Richtung 7) zugeordnet werden. Die Bedingung, um in Richtung 6) eingeordnet zu werden ist leer. Es ist unklar, wie es zu diesem marginalen Fehler gekommen ist, im definitiven Baum tritt er jedenfalls nicht mehr auf.

Obwohl schwer zu lesen, da es bei den Labels zum sog. Overplotting kommt, ist das zugehörige Dendrogramm in der folgenden Abbildung dargestellt. Entgegen den Bezeichnungen im Datensatz (wie auch in der Tabelle) sind die Faktorenniveaus durch Buchstaben und nicht durch Zahlen gekennzeichnet. Die vertikale Länge der Äste einer Verzweigung widerspiegelt die Höhe der Devianz und damit der «Wichtigkeit» dieser Verzweigung für die Erklärungskraft.

Die Verzweigung ist so zu lesen, dass wenn das Kriterium erfüllt ist, geht es links im Baum weiter, falls nicht, rechts.

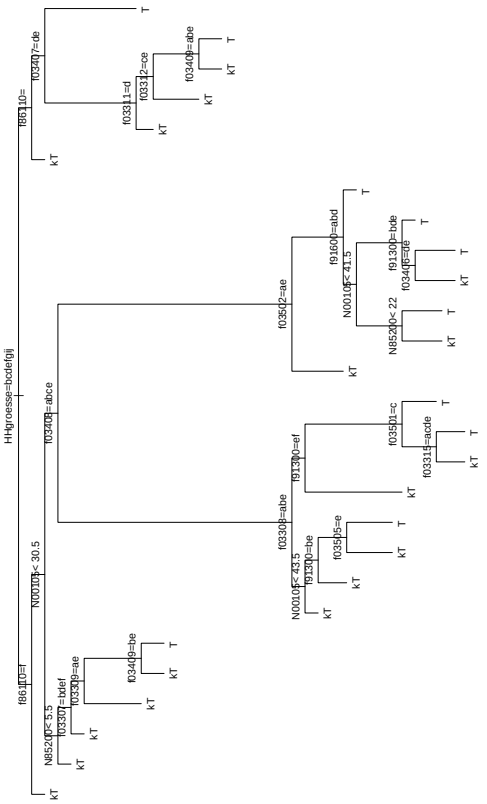


Abbildung 7.1: Dendrogramm mit allen Variablen

7.2 Suche nach optimalem Baum

Zur Bestimmung des optimalen Baums können zwei Schritte unterschieden werden. Zunächst muss eine Auswahl der zu verwenden Variablen getroffen werden, in einem zweiten Schritt muss die optimale Baumtiefe bestimmt werden.

Um eine Auswahl der Variablen zu treffen, wurde zunächst «manuell» vorgegangen. Es ist klar, dass demografische Fragen einschliesslich der Fragen zum Internetverhalten von LINK beim Rekrutierngsinterview prinzipiell gestellt werden. Zusätzlich stehen fünf weitere mögliche Fragen zur Auswahl. Um diese weiteren Variablen zu bestimmen, wurden die demografischen Variablen jeweils mit den Variablen aus den Fragekomplexen Big Five, Schwartz-Skala und Rational Choice kombiniert. Diejenigen Variablen, die eine Prognosekraft hatten wurden in einem Zwischenmodell zusammengefasst. Insgesamt handelt es sich dabei um zehn Variablen. Aus diesem Zwischenmodell konnten durch Ausprobieren ein Modell bestimmt werden, dass neben den demografischen nur fünf weitere umfasst. Ausprobieren bedeutet hier, dass jeweils eine Variable probeweise weggelassen wurde. Sinkt die Rate der Fehlklassifikationen nur gering, wurde auf diese Variable ganz verzichtet.

Die letztendlich verwendeten Variablen wurden im definitiven Modell zusammengefasst.

Da bei Baum-Modellen auch zum Teil hochdimensionale Wechselwirkungen zwischen Variablen berücksichtigt werden, ist eine solche manuelle Auswahl problematisch, da nicht abzusehen ist, wie eine Variable im Kontext anderer Variablen im Modell reagiert. Technisch ist es nur durch das Ausprobieren aller Kombinationen möglich, einen optimalen Baum zu bestimmen. Durch die begrenzte zur Verfügung stehende Rechenleistung ist dies allerdings unmöglich: Insgesamt stehen 33 Variablen zur Verfügung, die auf die 5 Plätze verteilt werden können. Es ergeben sich daher $33!/(33-5) = 28'480'320$ Bäume. Selbst wenn man alle ausprobiert, müssten die jeweils entstandenen Bäume noch einer Vergleichsprüfung unterzogen werden. Es wurde daher der Baum, der sich aus der manuellen Auswahl ergibt mit einer Stichprobe von 10'000 möglicher anderer Bäume verglichen.

Die folgende Abbildung zeigt die Verteilung der Fehlklassifikationsraten

für eine Stichprobe von 10'000 Bäumen:

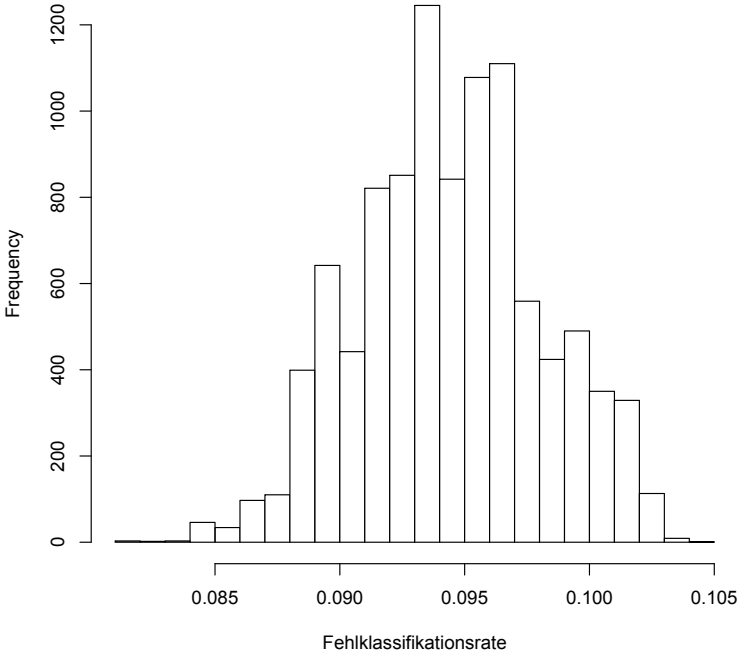


Abbildung 7.2: Verteilung der Fehlklassifikationsraten bei 10'000 zufällig ausgewählten Bäumen

Die Fehlklassifikationsrate des definitiven Baums liegt mit 0.083 nur leicht über dem Minimum der Stichprobe der 10'000 Bäume, welches 0.082 beträgt. Berücksichtigt man die fehlende Stutzung des Baums ist die Fehlklassifikationsrate des definitiven Baums in einem akzeptabel niedrigem Bereich.

7.3 Definitiver Baum

Die in der folgenden Auflistung aller verwendeten Variablen mit einem «†» gekennzeichnete Variable nach dem Zivilstand wird für die Entwicklung des eigentlichen Baums nur indirekt verwendet. Das heisst, dass sie nur wichtig ist, um die fehlenden Werte zu zählen, selbst aber nicht verwendet wird.

Demografische Fragen	
f90300†	Was ist Ihr Zivilstand?
N00105	Darf ich fragen wie alt Sie sind (Kategorisiert)
N85200	Wie oft haben Sie in den letzten 12 Monaten etwas online im Internet gekauft?
N90405	Anzahl Erwachsene 20-64 Jahre
f91100	Welche Schule haben Sie zuletzt besucht?
f91300	Persönliches Einkommen (6er Skala)
f00110	Geschlecht der befragten Person
Schwartz Skala	
f03405	. . .sich immer an Regeln zu halten
f03409	. . .sich immer richtig zu verhalten
Big Five Skala	
f03307	. . .wo eher faul ist
Rational Choice Fragen	
f03502	Umfragen bringen Abwechslung und sind interessant.
f85100	Wie häufig etwa nutzen Sie das Internet für private Zwecke?
Sonstiges	
noNA	Anzahl fehlender Werte bei allen verwendeten Variablen.

Insgesamt wurden 172 Befragte falsch klassifiziert, das entspricht 8.3 %. Im Vergleich zum Baum, für den alle Variablen verwendet wurden, sind dies nur 4 Befragte mehr, der Informationsverlust ist daher zu vernachlässigen. Der entstandene Baum hat folgende Knoten:

Fortsetzung nächste Seite . . .

Tabelle 7.2: Knoten des definitiven Baums

Knoten	Bedingung	Split	Anzahl	Devianz	Teilnahme	p(kT)	p(T)	

Tabelle 7.2: Knoten des definitiven Baums

Knoten	Bedingung	Split	Anzahl	Devianz	Teilnahme	p(kT)	p(T)	
1	root		2068	316	kT	0.85	0.15	
2	noNA	>=2.5	1281	7	kT	0.99	0.01	*
3	noNA	<2.5	787	309	kT	0.61	0.39	
6	N00105	<29.5	276	57	kT	0.79	0.21	
12	N85200	<5.5	206	31	kT	0.85	0.15	*
13	N85200	>=5.5	70	26	kT	0.63	0.37	
26	f03307	=2,4,5,99	34	6	kT	0.82	0.18	*
27	f03307	=1,3	36	16	T	0.44	0.56	
54	f91100	=3	14	2	kT	0.86	0.14	*
55	f91100	=2,4,5,6	22	4	T	0.18	0.82	*
7	N00105	>=29.5	511	252	kT	0.51	0.49	
14	noNA	>=0.5	179	65	kT	0.64	0.36	
28	N90405	=0,2,3,4	80	16	kT	0.8	0.2	*
29	N90405	=1	99	49	kT	0.51	0.49	
58	noNA	>=1.5	16	1	kT	0.94	0.06	*
59	noNA	<1.5	83	35	T	0.42	0.58	
118	f03409	=2,4,5	53	23	kT	0.57	0.43	
236	f91300	=2	17	3	kT	0.82	0.18	*
237	f91300	=1,3,4,5,6	36	16	T	0.44	0.56	
474	f00110	=2	12	2	kT	0.83	0.17	*
475	f00110	=1	24	6	T	0.25	0.75	*
119	f03409	=1,3	30	5	T	0.17	0.83	*
15	noNA	<0.5	332	145	T	0.44	0.56	
30	f03502	=1,3,5	180	83	kT	0.54	0.46	
60	f03405	=1,4,5	128	51	kT	0.6	0.4	
120	f03307	=4,5	13	0	kT	1	0	*
121	f03307	=1,2,3	115	51	kT	0.56	0.44	
242	f91300	=2,5	26	6	kT	0.77	0.23	*
243	f91300	=1,3,4,6	89	44	T	0.49	0.51	
486	N85200	<2.5	35	11	kT	0.69	0.31	*
487	N85200	>=2.5	54	20	T	0.37	0.63	*
61	f03405	=2,3	52	20	T	0.38	0.62	
122	f03409	=1,3,5	29	12	kT	0.59	0.41	*
123	f03409	=2,4	23	3	T	0.13	0.87	*
31	f03502	=2,4	152	48	T	0.32	0.68	
62	N00105	<43.5	72	32	T	0.44	0.56	
124	f03502	=4	35	14	kT	0.6	0.4	
248	f85100	=3,5,6	27	8	kT	0.7	0.3	*

Fortsetzung nächste Seite . . .

Tabelle 7.2: Knoten des definitiven Baums

Knoten	Bedingung	Split	Anzahl	Devianz	Teilnahme	p(kT)	p(T)	
249	f85100	=4	8	2	T	0.25	0.75	*
125	f03502	=2	37	11	T	0.3	0.7	*
63	N00105	>=43.5	80	16	T	0.2	0.8	*

* bezeichnet Endknoten

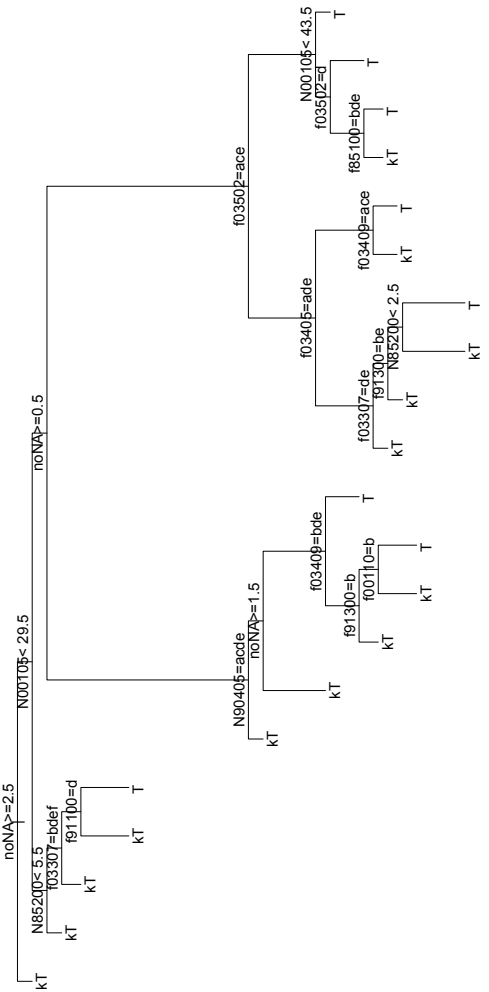


Abbildung 7.3: Dendrogramm des definitiven Baums

Der Baum entspricht einem «gestutzten» Baum (*pruned tree*). Bei der Entwicklung von Bäumen kann entschieden werden, welche Tiefe die Bäume haben dürfen, d. h. wie viele Verästelungen unterhalb des Ursprungsknotens (*root*) erlaubt werden. Es ist notwendig Bäume zu trimmen, um ein *over-fitting* zu vermeiden, d. h. es wird nicht mehr die Struktur der Zusammenhänge abgebildet, sondern das Modell / der Baum wird soweit verfeinert, dass auch unnötigerweise Fehler (*noise*) erfasst werden. Zu erkennen ist dies daran, dass die resultierenden Zellen der Blätter des Baums nur sehr wenige Fälle enthalten.

Um zu überprüfen, ob der Baum tatsächlich das Antwortverhalten abbilden kann und nicht nur overfitted ist, wurde eine Vergleichsprüfung (*cross-validation*) vorgenommen. Aus der Vergleichsprüfung ergeben sich Missklassifikationsraten, die von der des eigentlichen Baums abweichen und über dieser liegen.

7.3.1 Vergleichsprüfung und Stutzen des Baums

Ein Mass für die Güte eines Baumes ist die durchschnittliche Fehlkalssifikationsrate für alle n Blöcke. Es ermöglicht zu untersuchen, wie «tief» ein Baum werden darf, ohne nur Zufallseffekte zu erfassen. Um einen optimalen Wert zu ermitteln, der hilft festzulegen, wie weit der Baum gestutzt werden muss, wie verästelt er also sein darf, können verschiedene Werte ausprobiert werden und die entstandenen gestutzten Bäume mittels Vergleichsprüfung verglichen werden.

Die folgende Abbildung zeigt die durchschnittliche Fehlklassifikation der Vergleichsprüfung der gestutzten Bäume, abhängig vom die Tiefe der Bäume bestimmenden Kriterium *cp*.

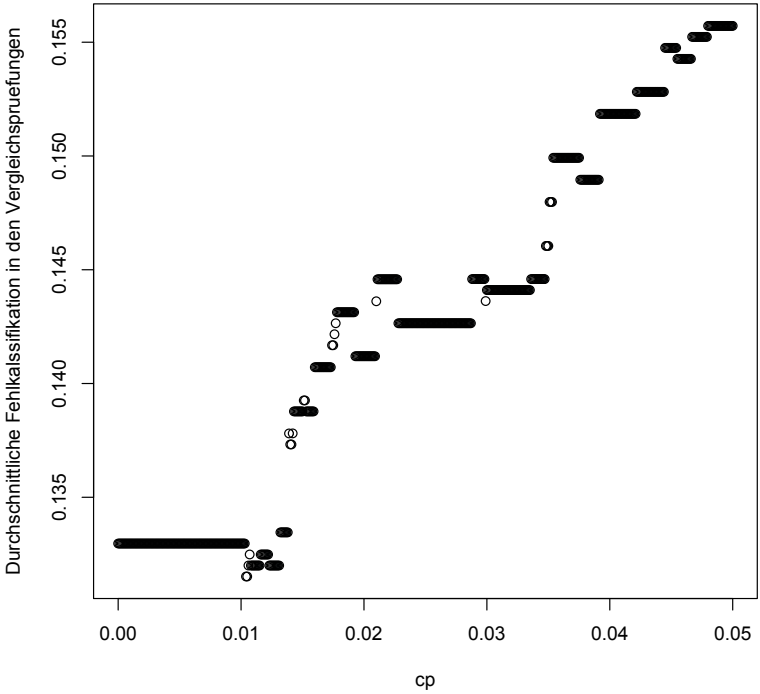


Abbildung 7.4: Vergleichsprüfung

Der optimale Wert für *cp* liegt zwischen 0.0104 und 0.0105, hier ist also die Fehlklassifikationsrate am geringsten. Zwischen diesen beiden Werten beträgt die durchschnittliche Fehlklassifikationsrate der Vergleichsprüfungen konstant 13.1518%. (Die Abbildung zeigt nur den Bereich für *cp*s zwischen 0 und 0.05. Wird das *cp* grösser gewählt, nimmt die Fehlklassifikationsrate nur noch ab, da der Baum dann zu sehr gestutzt wird.)

7.3.2 Resultierende Teilnahmewahrscheinlichkeiten

Die folgende Tabelle 7.3 listet alle entstandenen Teilnahmewahrscheinlichkeiten auf.

Tabelle 7.3: Teilnahmewahrscheinlichkeiten gemäss Baum

Teilnahme-wahrscheinlichkeit	Anzahl in GG	Anzahl in S
0.00	13	0
0.66	1'060	7
6.25	16	1
14.29	14	2
15.05	206	31
16.67	12	2
17.65	51	9
20.00	80	16
23.08	26	6
29.63	27	8
31.43	35	11
41.38	29	12
62.96	54	34
70.27	37	26
75.00	32	24
80.00	80	64
81.82	22	18
83.33	30	25
86.96	23	20

Ein hoher Anteil der telefonisch Befragten haben eine sehr geringe Wahrscheinlichkeit in die Stichprobe zu gelangen. Rund 58 % der 1'847 Befragten haben eine Wahrscheinlichkeit von unter einem einem Prozent am Web-Panel teilzunehmen. Diese Gruppe hat dann auch einen entsprechend kleinen Anteil von 2.2 % bei den Freiwilligen. Über den grössten Teil der angestrebten Grundgesamtheit liegen also im Web-Panel keine Informationen vor.

Insofern ist es fraglich, ob eine Gewichtung tatsächlich eine Verbesse-

rung der Datenqualität ermöglichen kann. Vereinfacht gesprochen muss folgendes Kriterium erfüllt sein, um eine Bias-Reduktion im Fall von Missing-at-random zu ermöglichen: Die interessierende Variable Y muss mit der modellierten Teilnahmewarscheinlichkeit korrelieren.

Im vorliegenden Fall unterscheiden sich diejenigen, die am Panel teilnehmen wollen, kaum von denjenigen, die dies verweigern. Insbesondere gibt es keine Unterschiede bei den Variablen, die als Kontrollvariablen in die Befragung implementiert wurden (Präferenz für einen Supermarkt und Besitzt eines SBB-Abonements.)

Bei folgenden Variablen gibt es einen Unterschied gemäss χ^2-Test (5 % Sig. Niveau):

`f85100`	Wie häufig benutzen Sie das Internet für private Zwecke?
`f03505`	Wichtigkeit sich an Regeln zu halten
`f03404`	Wichtigkeit im Leben Abwechslung zu haben
`f03501`	Marktforscher behandeln die Daten vertraulich
`f03502`	Umfragen sind abwechslungsreich und interessant
`f91200`	Haushaltseinkommen

Es handelt sich also nur um sehr wenige Variablen, bei denen ein Unterschied festzustellen ist. Mit anderen Worten, der Ausfallmechanismus bewirkt in vielen Fällen keinen Bias bei Y. Dieses Ergebnis ist jedoch nicht verallgemeinerbar, d. h. es ist nicht ad hoc klar, dass der Nonresponse unproblematisch ist.

An dieser Stelle muss nochmals angemerkt werden, dass der Vergleich derjenigen, die bereit sind am Panel teilzunehmen nur mit denjenigen erfolgen kann, die an der CATI-Befragung partizipiert haben. Der Unterscheid zwischen beiden Gruppen scheint nicht sehr gross zu sein, wenngleich jedoch existent. Das ist insofern überraschend, als dass Verhalten bei der Befragung doch sehr unterschiedlich gewesen ist. Es ist z. B. so, dass die Freiwilligen im Gegensatz zu den Verweigerern sehr viel weniger fehlende Werte bei den Antworten produziert haben. Die Verweigerungshaltung ist also schon im CATI spürbar gewesen.

Auch bei den später zu beschreibenden Modellierungen des Antwortverhaltens zeigt sich, dass die Zahl fehlender Antworten die Teilnahmewahrscheinlichkeit am besten prognostizieren kann. Man kann daher vermuten, dass diejenigen mit vielen fehlenden Werten den Totalverweigerern (als

denjenigen, die schon das CATI verweigert haben) am ähnlichsten sind. Aus theoretischen Überlegungen ist diese Personengruppe also besonders wichtig, wenn es darum geht, von einer Web-Befragung auf die gesamte Bevölkerung schlussfolgern zu wollen. Trivialerweise ist es aber schwierig ohne Angaben das Responseverhalten zu modellieren.

Als Empfehlung an die Rekrutierung gilt es festzuhalten, dass es eine deutliche Verbesserung der Datenqualität dadurch erreicht werden kann, den Anteil der Verweigerer möglichst klein zu halten. mit rund 110'000 Panelisten ist das LINK-Panel sehr gross und auch geeignet, kleine Subpopulationen abzubilden. Es empfiehlt sich daher zukünftig weniger Aufwand darin zu investieren, möglichst viele weitere Panelisten zu rekrutieren, sondern sicherzustellen, dass das Rekrutierungsinterview besonders motivierend ist. Möglicherweise hilft es dabei, verscheidende Rekrutierungsstile zu vergleichen.

Empfiehlt es sich beispielsweise, die Rekrutierungsfrage an laufende Telefonbefragungen anzuhängen oder sollte die Rekrutierung lieber durch ein separates Interview realisiert werden? Weiterhin sollte überprüft werden, ob es nicht geschickt ist, eine Spezialisierung des Interviewerpersonals zuzulassen, um so beispielsweise eine Investition in gezielte Fördermassnahmen beispielsweise durch Psychologen zu rechtfertigen.

Die Annahme, dass diejenigen mit den hohen Teilnahmewahrscheinlichkeiten den Teilnehmern ähnlicher sind als die mit kleinen Teilnahmewahrscheinlichkeiten, stimmt allerdings nicht. Die CATI-Befragten wurden versuchsweise in drei Gruppen eingeteilt: die grosse Gruppe derer mit sehr niedrigen Teilnahmewahrscheinlichkeit von unter einem Prozent, den sonstigen nicht-Teilnehmern und den Panelisten. Es zeigt sich, dass sich die Panelisten in den selben Variablen und nur in diesen von denjenigen mit niedrigsten Teilnahmewahrscheinlichkeiten unterscheiden, wie auch von denjenigen mit höherer Teilnahmewahrscheinlichkeit, die trotzdem nicht teilgenommen haben. (Diesen Umstand kann man auch so interpretieren, dass die Modellierung der Teilnahmewahrscheinlichkeit vermutlich noch nicht sehr gut ist.)

Es ist allerdings möglich, einen Vergleich zwischen den beiden Gruppen bezüglich extern bekannter Merkmale anzustellen:

Als Fazit bleibt festzuhalten, dass da die Unterschiede zwischen den beiden Gruppen nur sehr klein ist, es schwierig wird zu überprüfen, inwiefern das

Gewichtungsverfahren zu einer Verbesserung der Datenqualität führen kann.

8 Auswahl der Kovariaten mittels Probit Modellierung

► Eine sehr verbreitete Methode der Modellierung sind Lineare Modelle. Sie werden sehr häufig angewendet, nicht nur, weil sie in allen gängigen Softwarepacketen für statistische Analysen standardmässig implementiert sind, sondern auch, weil es viele Erfahrungen mit der Einschätzung der Güte eines solchen Modells gibt. Weil die Verbreitung so gross ist, soll auch hier die Teilnahmewahrscheinlichkeit nochmals mittels eines Linearen Modells modelliert werden. Das dient dem Vergleich mit dem Baum-Modell. Ausserdem wird in den sich anschliessenden Kapitel versucht, das Verhalten der *propensity scores* noch besser mittels Simulationen zu verstehen, wozu auch auf Lineare Modellierung zurückgegriffen werden wird. ◄

Um den Einfluss der erhobenen Variablen auf die Bereitschaft der Teilnahme alternativ zu einem Baum zu schätzen, wurde ein Probitmodell bestimmt. Probitmodelle sind eine Spezifikation des Generalisierten Linearen Modells (GLM) mit einer Probit-Link Komponente. Ein Probitmodell hat zur Annahme, dass gilt

$$Pr(Y = 1|X = x) = \Phi(x'\beta)$$

wobei Y den Response (Teilnahme / Verweigerung) an der ersten Online-Befragung, X die Kovariaten und Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. β kann mit Hilfe einer Maximum Likelihood Methode geschätzt werden, Collett (2002)

Auswahlkriterien

Die Suche nach einem möglichst guten Modell wurde mit Hilfe verschiedener Kriterien vorgenommen. Folgende Kriterien wurden verwendet, um die Modellreduktion und dann -anpassung vorzunehmen[1]:

1 Für eine Einführung siehe z. B. Burnham und Anderson (2004).

Akaike Information Criterion (AIC) Das AIC berechnet sich als $-2 \log \mathcal{L} + 2npar$, wobei $npar$ der Anzahl Parameter im Modell entspricht und $\mathcal{L}$ die Likelihood ist. Der erste Summand wird auch als Devianz D bezeichnet.

Bayes Information Criterion (BIC) auch bekannt als **Schwartz Bayes Criterion (SBC)** Das BIC unterscheidet sich dadurch, dass npar nicht mit 2 sondern mit $\log n$ multipliziert wird. Ein Modell ist um so besser, je kleiner das AIC bzw. das SBC ist, wobei letzteres die Anzahl verwendeter Parameter stärker «bestraft» und zwar abhängig von der Anzahl verwendeter Fälle.

Pseudo-Bestimmtheitsmasse Bei Maximum-Likelihood Schätzungen ist es nicht möglich, das Bestimmtheitsmass R^2 (auch B abgekürzt) als Quadrat des multiplen Korrelationskoeffizienten zu berechnen. In solchen Modellen werden sogenannte Pseudo-Bestimmtheitsmasse verwendet. Aus der Liste möglicher Bestimmtheitsmasse sollen hier die drei populärsten verwendet werden, die wie folgt definiert sind (Kutner et al., 2004):

McFadden $R^2_M = 1 - D/D_0$, wobei D_0 die Schätzung für die mittlere Devianz in der Population ist, wenn das Modell lediglich aus einer Konstanten besteht.

Cox-Snell $R^2_C = 1 - \exp(D_0 - D)$.

Nagelkerke $R^2_N = \frac{1-\exp(D_0-D)}{1-\exp(-D_0)}$.

ANOVA Der Einfluss der einzelnen Variablen kann mit Hilfe einer Varianzanlyse getestet werden. Da GLMs mit der Maximum Likelihood Methode geschätzt werden, handelt es sich um eine Analys der Deviance. In R ist dies als Funktion «anova()» implementiert, die einen Likelihood-Quotienten-Test durchführen kann.(Baily, 2008)

Multikollinearität Die Kollinearität von Variablen kann mittels κ (Kappa) gemessen werden. κ ist im Fall eines linearen Modells definiert als

$$\kappa = \frac{\sqrt{\lambda_{\max}}}{\sqrt{\lambda_{\min}}},$$

wobei $\lambda_{\max}$ der grösste Eigenwert der Matrix $X'X$ ist und $\lambda_{\min}$ der kleinste. (X ist die Design-Matrix des Modells).

Varianz-Vergrösserungs-Faktor (variance inflation factor, VIF) Der VIF einer Variable beschreibt die Kollinearität dieser Variable mit allen anderen Variablen: VIF= $1/(1-R_j^2)$, wobei R_j^2 das Bestimmtheitsmass einer Regression von x_j auf alle anderen Variablen im Modell ist.

Binned Plot Der Binned Plot ist ein Plot von mittleren Residuen gegen mittlere angepasste Werte. Die Mittelung ist notwendig, weil die rohen Residuen eines logistischen Modells nur die Werte 1 oder −1 annehmen können. Die Anzahl «Bins» (Sammelabschnitte) für die Mittelung wurde entsprechend dem Weglasswert der Funktion `binnedplot` im R Package «arm» gewählt ($\lfloor\sqrt{n}\rfloor$). Die grauen Linien geben den Bereich ±2 mal Standartfehler an, in dem 95% aller Werte des Bins liegen sollten.

Normal-Verteilungs-Diagramm (Quantile-Quantile-Normal Plot, Q-Q-Normal Plot) Es ist hilfreich, die Verteilung der Devianz-Residuen zu betrachten. Devianz-Residuen werden berechnet als $sign(y_i - \hat{y}_i)/\sqrt{d_i}$, wobei d_i die i-te Komponente der Devianz ist.

Rückwärts-Elimination

Es wurden verschiedene Methoden ausprobiert, ein möglichst gutes Modell zu finden. Ausgehend von einem Referenzmodell mit allen erhobenen Variablen, konnte anhand von Informationskriterien und geeigneten Grafiken eine Modellreduzierung durch Rückwärts-Elimination vorgenommen werden, um sowohl einen guten Modell-Fit zu erreichen, wie auch um eine Reduktion der Variablen vornehmen zu können.

Nachdem mittels Rückwärts-Elimination ein kleines stabiles Modell gefunden wurde, wurden dem Modell wieder versuchsweise alle Variablen einzeln hinzugefügt. Dabei wurde auch versucht, die Erklärungskraft der Variablen gegebenenfalls mittels geeigneter Umkodierungen zu verbessern.

Die Kodierung der Daten wurde bei den Probit-Modellen beibehalten. Insbesondere gilt dies auch für die Zuweisung, ob es sich um eine stetige oder kategoriale Variable handelt. Selbstverständlich wurden die Variablen denen eine Likert-Skala zugrunde liegt auch versuchsweise als quasi-metrisch angenommen. Das hat aber in keinem der Versuche zu einer Modellverbesserung geführt, im Gegenteil. Wie auch die Ergebnisse aus den Baum-Modellen

gezeigt haben, ist es häufig nicht richtig, Likert-Skalen als stetig zu interpretieren. Der Übersichtlichkeit halber wurde auf die Darstellung der Modelle verzichtet, die versuchsweise andere Kodierungen als die bei den Bäumen verwenden.

Das bedeutet auch, dass Befragte mit fehlenden Werten nicht ausgeschlossen werden mussten, da die Information «fehlender Wert» bei allen kategorialen Variablen eine Kategorie neben den anderen Kategorien geblieben ist. Da es sich bei fast allen Variablen um kategoriale handelt, mussten keine Befragten ausgeschlossen werden, was aber auch zu einer Erhöhung der Multikolinearitäten führt.

Gelegentlich ist es sogar inhaltlich sehr viel besser, eigentlich stetige Variablen als kategorial zu behandeln, wie z. B. beim Einkommen. Das Einkommen wurde von LINK kategorisieret erfasst, wobei man auch diese Kategorisierung eigentlich als quasi-metrisch interpretieren könnte. Argumentiert man aber inhaltlich, dass die Teilnahmewahrscheinlichkeit nicht bei Personen mit entweder besonderes hohen oder niedrigen Einkommen besonders hoch ist, sondern dass es im Gegenteil einen Mittelschichtbias gibt, kann dieser in Regressionen am besten erfasst werden, wenn man das sowieso schon kategorisierte Einkommen auch als kategoriale Variable behandelt.

8.1 Diskussion des Referenzmodells

Im Referenzmodell wurden alle erhobenen Variablen als erklärend verwendet, keine dieser variablen wurde umkodiert. Es ergibt sich ein Modell, welchem trotz relativer hoher Pseudo-Bestimmtheitsmasse nicht zu trauen ist, da es nicht nur zu perfekten Vorhersagen kommt, sondern insbesondere die Multikollinearitäten sehr hoch sind.

	McF.R^2	CSR^2	Nag.R^2	AIC	BIC/SBC	κ
Referenzmodell	0.58	0.39	0.68	1'202	1'979	*Inf*

Tabelle 8.1: Gütekriterien des Referenzmodells

Da die Koeffizienten des Modells in diesem Zusammenhang nicht sehr interessant sind, folgt nur die Tabelle der ANOVA:

Tabelle 8.2: ANOVA Referenzmodell

	Df	Deviance	Resid. Df	Resid. Dev	P(> $\|\chi\|$)	
NULL			2067	1768.32		
N00105	1	1.06	2066	1767.26	0.3030	
f00110	1	2.38	2065	1764.88	0.1228	
f00140	2	21.91	2063	1742.97	0.0000	***
f03200	5	8.85	2058	1734.12	0.1151	
f03301	5	11.12	2053	1722.99	0.0490	*
f03302	5	8.05	2048	1714.94	0.1533	
f03303	5	4.23	2043	1710.71	0.5165	
f03304	5	8.00	2038	1702.71	0.1563	
f03305	5	5.38	2033	1697.33	0.3715	
f03306	5	1.42	2028	1695.91	0.9216	
f03307	5	5.35	2023	1690.56	0.3751	
f03308	5	14.31	2018	1676.25	0.0138	*
f03309	5	12.09	2013	1664.17	0.0336	*
f03310	5	4.62	2008	1659.55	0.4640	
f03311	5	4.43	2003	1655.12	0.4892	
f03312	5	2.01	1998	1653.11	0.8478	
f03313	5	5.38	1993	1647.72	0.3710	
f03314	5	6.61	1988	1641.12	0.2515	
f03315	5	3.20	1983	1637.92	0.6695	
f03401	5	2.57	1978	1635.34	0.7653	
f03402	5	5.15	1973	1630.19	0.3977	
f03403	5	3.73	1968	1626.46	0.5889	
f03404	5	4.69	1963	1621.78	0.4555	
f03405	5	17.59	1958	1604.18	0.0035	**
f03406	5	2.73	1953	1601.45	0.7414	
f03407	5	7.21	1948	1594.25	0.2057	
f03408	5	6.94	1943	1587.31	0.2255	
f03409	5	2.94	1938	1584.37	0.7095	
f03410	5	10.04	1933	1574.33	0.0742	.
f03501	5	15.72	1928	1558.62	0.0077	**

Fortsetzung nächste Seite . . .

Tabelle 8.2: ANOVA Referenzmodell

	Df	Deviance	Resid. Df	Resid. Dev	P(> |χ|)	
f03502	4	23.60	1924	1535.01	0.0001	***
f03503	4	2.48	1920	1532.53	0.6477	
f03504	4	12.59	1916	1519.94	0.0135	*
f03505	4	11.15	1912	1508.79	0.0249	*
f85100	5	90.98	1907	1417.81	0.0000	***
N85200	1	7.98	1906	1409.83	0.0047	**
f85300	3	6.58	1903	1403.25	0.0864	.
f86110	6	579.70	1897	823.55	0.0000	***
f90300	4	21.28	1893	802.27	0.0003	***
N90401	3	0.68	1890	801.58	0.8772	
N90402	3	0.24	1887	801.35	0.9713	
N90403	4	3.11	1883	798.24	0.5400	
N90404	3	3.39	1880	794.85	0.3354	
N90405	8	21.64	1872	773.21	0.0056	**
N90406	3	2.65	1869	770.56	0.4486	
f91100	7	4.56	1862	765.99	0.7131	
f91300	6	16.77	1856	749.22	0.0102	*
f91600	4	1.34	1852	747.88	0.8549	
noNA	0	0.00	1852	747.88	—	

Signif. Niveau: 0 *** 0.001 ** 0.01 * 0.05 .

Aus Sicht des Referenzmodells sind die wichtigsten Variablen, das Teilnahmeverhalten zu erklären, der Zivilstand und die Fragen aus der Gruppe der Rational Choice Fragen. Der Binned Plot der Residuen in Abb, 8.1 zeigt, dass die Verteilung der Residuen nicht unproblematisch ist. Es scheint jedenfalls so zu sein, dass die Residuen mit zunehmender Wahrscheinlichkeit teilzunehmen immer kleiner werden.

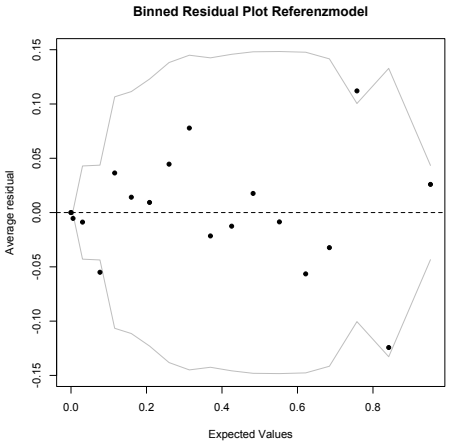


Abbildung 8.1: Binned Plot der Residuen im Referenzmodell

Ein wichtiges Kriterium zur Beurteilung des Modells ist es, zu überprüfen wie gut es die Teilnahme an der Befragung prognostizieren kann. Als Teilnehmer im Sinne des Modells gilt, wer eine höhere angepasste Werte als 0.5 hat. Tabelle 8.3 vergleicht die prognostizierte Teilnahme des Baums mit der des Referenzmodells. Vergleich Baum

		Baum		Referenzmodell	
	korrekt	kT	T	kT	T
kT	1'752	1'685	67	1'672	80
T	316	105	211	114	202

T: Teilnahme, kT: keine Teilnahme

Tabelle 8.3: Prognose der Teilnahme im Referenzmodell

Die Unterschiede zwischen beiden Modellen sind nicht gross. Wichtiger noch als die absolute Rate der Fehlklassifikationen ist dabe der Anteil an Teilnehmern, die vom jeweiligen Modell richtig erkannt werden. Dieser Anteil beträgt beim Baum $211/316 = 66.77\%$ und liegt damit leicht über

dem Anteil des Probitreferenzmodell mit 63.92 %. Selbst im Vergleich zu einem Probit-Modell, welches alle Variablen enthält, schneidet ein Baum mit reduzierter Anzahl Variablen besser ab.

8.2 Zwischenmodell I mit imputierten fehlenden Werten

Um auszuschliessen, dass ein Probit-Modell möglicherweise doch eine bessere Vorhersagekraft hat, als ein Baum, wurden ein Kandidat für ein definitives Modell unter Ausschluss der Beobachtungen bestimmt, die fehlende Werte aufweisen. Somit konnten diejenigen Variablen identifizieret werden, die unabhängig von den fehlenden Werten eine möglichst hohe Prognosekraft haben. In einem zweiten Schritt wurde dann für die fehlenden Werte der Variablen dieses Modells jeweils ein Imputationsmodell entwickelt, um so das ursprünglich entwickelte Modell auf alle Beobachtungen anwenden zu können.

Es ist klar, dass dies kein praktikables Vorgehen im Alltag von Befragungen ist, den *propensity score* zu bestimmen. Der Aufwand wäre jeweils einfach zu hoch. Nichtsdestoweniger soll ausgeschlossen werden, dass nicht möglicherweise doch wichtige Informationen verloren gehen. Ausserdem liefert ein solches Modell eine gute Vergleichsmöglichkeit für weitere Modelle.

Das hier entwickelte Zwischenmodell verzichtet auf die Berücksichtigung der Beschränkung auf fünf Variablen plus demographische Variablen, um ein möglichst gute Modell zu entwickeln.

1'454 Beobachtungen konnten berücksichtigt werden. Das Zwischenmodell enthält folgende Variablen: `f00140`, `f03308`, `f03502`, `f85100`, `f90300` und `noNA`. Tabelle 8.4 zeigt die Koeffizienten.

Vergleich Referenzmodell

Die Informationskriterien des Zwischenmodell (Tabelle 8.5) haben sich im Vergleich zum Referenzmodell bezüglich der Pseudo-Bestimmtheitsmasse etwas verschlechtert. Da die Länge der Datensätze unterschiedlich ist, können AIC und BIC nicht verglichen werden. Da die Mulitkolinearitäten weiterhin sehr hoch sind und es immer noch zu perfekten Vorhersagen aufgrund fehlender Werte kommt, ist selbst das Ablesen einer Tendenz bei den Kriterien schwierig.

Vergleich mit komplettem Datensatz

Wendet man das Modell wiederum auf den gesamten Datensatz an, ohne zu berücksichtigen, ob eine Beobachtung fehlende Werte enthält, ergibt sich eine vergleichsweise schlechte Prognose der Teilnahmebereitschaft, siehe

	Df	Deviance	Resid. Df	Resid. Dev	P(> $\|\chi\|$)	
NULL			1453	1504.23		
f00140	2	25.59	1451	1478.64	0.0000	***
f03308	5	23.01	1446	1455.62	0.0003	***
f03502	4	28.22	1442	1427.40	0.0000	***
f85100	5	103.00	1437	1324.40	0.0000	***
f90300	3	23.78	1434	1300.62	0.0000	***
noNA	1	301.80	1433	998.82	0.0000	***
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .						

Tabelle 8.4: ANOVA Zwischenmodell I

	McF.R^2	CSR^2	Nag.R^2	AIC	BIC/SBC	κ
Referenzmodell	0.58	0.39	0.68	1'202	1'979	*Inf*
Zwischenmodell I	0.34	0.29	0.46	1'040	1'062	$\approx$ 100'000

Tabelle 8.5: Gütekriterien des Zwischenmodells I

Tabelle 8.6.

		Baum		Referenzmodell		Zwischenmodell I	
	korrekt	kT	T	kT	T	kT	T
kT	1'752	1'685	67	1'672	80	1'671	81
T	316	105	211	114	202	174	142
T: Teilnahme, kT: keine Teilnahme							

Tabelle 8.6: Prognose der Teilnahme im Zwischenmodell

Da auch dieses Modell wieder an den Beobachtungen mit den fehlenden Werten scheitert, ist versucht worden, diese zu imputieren. Die folgende Abbildung zeigt die Zahl fehlender Werte in den Variablen mit fehlenden Werten, für die eine Imputation vorgenommen werden soll.

Für die Analyse der fehlenden Werte ist die folgende Abbildung 8.2 Fehlende Werte

hilfreich[2]. Auf der linken Seite der Abbildung sind die Anteile der fehlenden Werte eingezeichnet. f90300 weist einen sehr hohen Anteil fehlender Werte auf, für f03308 und f03502 ist dieser moderat. Die rechte Seite ist ein Raster der Kombinationen von fehlenden Werten. Rote Felder stellen fehlende Werte dar, blaue Angaben. An der rechten Seite steht jeweils, wie hoch der Anteil der Befragten mit der entsprechenden Kombination fehlender Werte ist.

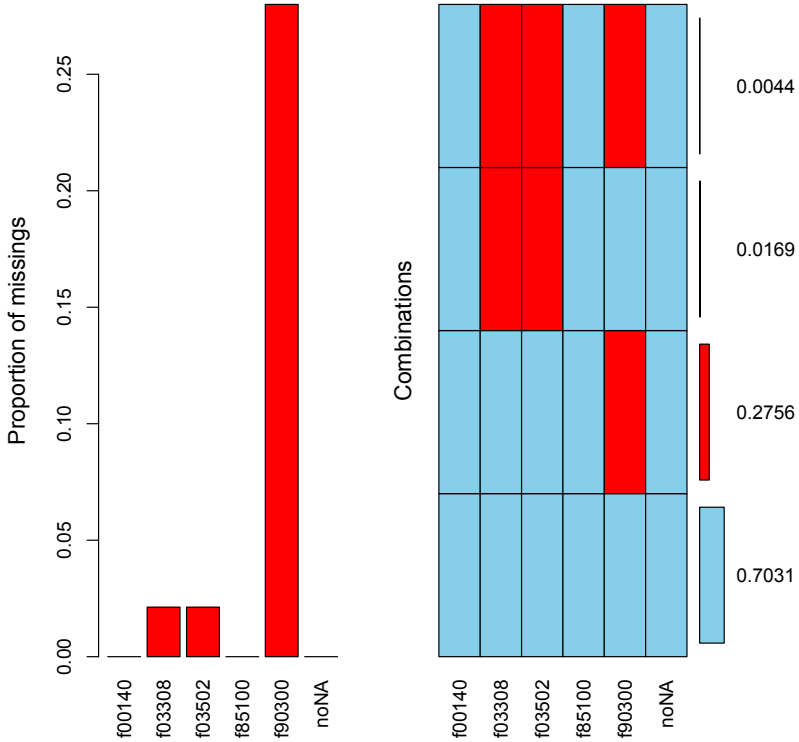


Abbildung 8.2: Fehlende Werte im definitiven Modell

2 Zur Erstellung der Grafik wurde das R Package VIM verwendet, Templ et al. (2010)

Variablen des Imputationsmodells

Für die Imputation soll darauf verzichtet werden, weitere Variablen als die des Zwischenmodells heranzuziehen, da im späteren Einsatz beim *propensity score* Adjustment in der Praxis auch keinen weiteren Variablen zur Verfügung stehen.

Es müssen stufenweise die fehlenden Werte einer Variable nach der anderen imputiert werden. Gemäss Abbildung 8.2 muss mit f03308 und f3502 begonnen werden, da der Anteil fehlender Werte bei diesen Variablen am kleinsten ist. Für die Impuation stehen damit zunächst nur drei Variablen zur Verfügung.

8.2.1 Imputation fehlender Werte bei `f03308`

Die Variable mit den nächst wenigen fehlenden Werten ist `f03308`. Da diese Variable kategorial im Sinne des definitiven Modells ist, muss eine Diskriminanzanalyse[3] vorgenommen werden (McLachlan, 2004). Für 49 Beobachtungen muss eine Imputation fehlender Werte durchgeführt werden.

Der Vergleich beider Verteilungen (Tabelle 8.7) zeigt, dass die prognostizierte Gruppenzuteilung auf die Faktorenstufen sehr unterschiedlich zwischen Prognose und tatsächlichen Werten ist.

	1	2	3	4	5
Prognose	0	0	33	15	1
in %	0	0	67	31	2
Tatsächliche Verteilung	43	147	464	772	593
in %	2	7	23	38	29

Tabelle 8.7: Prognose `f03308`

Die mittlere Gruppe 3 wird bei der Prognose deutlich bevorzugt, 1 und 2 werden nicht prognostiziert. Interpretativ ist es sehr schwierig einzuschätzen, inwieweit diese Prognose tatsächlich gerechtfertigt oder fehlerhaft ist. Versuchsweise wurde das Modell auf zufällig ausgewählte Befragte angewendet, wobei die Prognose jeweils auch die ersten beiden Antwortkategorien deutlich unterschätzt und Antwortkategorie 3 überschätzt hat.

3 Die Berechnung wurde in R mit Hilfe der Funktion `lda` aus dem Paket `MASS` durchgeführt (Venables und Ripley, 2002)

8.2.2 Imputation fehlender Werte bei f03502

Da auch diese Variable kategorial im Sinne des definitiven Modells ist, muss wiederum eine Diskriminanzanalyse vorgenommen werden, um 44 fehlende Werte zu imputieren.

Der Vergleich beider Verteilungen zeigt (Tabelle 8.8), dass die prognostizierte Gruppenzuteilung auf die Faktorenstufen sehr unterschiedlich zwischen Prognose und tatsächlichen Werten ist.

	1	2	3	4	5
Prognose	44	0	0	0	0
in %	100	0	0	0	0
Tatsächliche Verteilung	248	383	713	477	203
in %	12	19	35	24	10

Tabelle 8.8: Prognose f03502

Überraschenderweise wird die Antwortkategorie 1 für alle Befragten mit fehlenden Werten prognostiziert. Bei dem Versuch das Modell auf zufällig ausgewählte Beobachtungen ohne fehlende Werte zu übertragen war dem nicht so.

8.2.3 Imputation fehlender Werte bei f90300

Auch bei dieser Variable handelt es sich um eine kategoriale variable und entsprechend muss wieder eine Diskriminanzanalyse vorgenommen werden, um eine Imputation der fehlenden Werte vornehmen zu können. Es fehlen die Angaben von 579 Befragten.

Die Verteilung der prognostizierten Antworten ist in Tabelle 8.9 dargestellt.

	1	2	3	4
Prognose	256	323	0	0
in %	44	56	0	0
Tatsächliche Verteilung	577	720	43	149
in %	40	50	2	7

Tabelle 8.9: Prognose f03502

Auch bei dieser Prognose gilt wieder, dass die Verteilung der Antwortkategorien bei den prognostizierten Antworten gar nicht mit der tatsächlichen Verteilung übereinstimmt. Selbstverständlich ist eine solche Verteilung theoretisch möglich, da Befragte, die nicht auf eine Frage antworten vermutlich eine abweichende Verteilung haben. Dass aber kein Befragter, der nicht geantwortet hat, nicht doch zu den Antwortkategorien 3 oder 4 gehören.

Nur zur kurzen Illustration der Güte des Modells ein Vergleich der Verteilung der prognostizierten mit den tatsächlichen Antworten in Tabelle 8.10.

	1	2	3	4
Prognose	516	928	45	0
Tatsächliche Verteilung	577	720	43	149

Tabelle 8.10: Prognose `f03502` (alle Daten)

Antwortkategorie 4 wird vom Modell nicht prognostiziert. Das Modell der Diskriminanzanalyse ist offensichtlich nicht sehr gut: 771 der 1'489 (51 %) Befragten werden bezüglich ihrer Antwort auf Frage f90300 richtig klassifiziert. Hätte sich die Impuatation fehlender Werte innerhalb eines Probit-Models als wichtiges Element zur Prognose der Teilnahmebereitschaft erwiesen, müssten die drei Imputationsmodelle nochmals neu entwickelt werden.

8.3 Zwischenmodelle

8.3.1 Zwischenmodell I

Das Zwischenmodell I entspricht dem oben entwickelten Modell ohne fehlende Werte. Alle Variablen haben weiterhin einen hohen signifikanten Einfluss im Modell (Tabelle 8.11).

Das Modell mit imputierten fehlenden Werten hat bezüglich der Informationskriterien deutlich schlechter abgeschnitten als das Modell ohne Imputation, siehe Tabelle 8.12. Bezüglich der Rate der falsch klassifizierten Teilnahme ist das Modell geradezu unbrauchbar. Lediglich 35 der 316 Teilnehmer (11.1 %) werden von dem Modell richtig erkannt (Tabelle 8.13).

	Df	Deviance	Resid. Df	Resid. Dev	P(> |χ|)	
NULL			2067	1768.32		
f00140	2	17.76	2065	1750.56	0.0001	***
f03308	4	18.87	2061	1731.69	0.0008	***
f03502	4	31.15	2057	1700.54	0.0000	***
f85100	4	91.66	2053	1608.89	0.0000	***
f90300	3	30.74	2050	1578.15	0.0000	***
noNA	1	244.73	2049	1333.43	0.0000	***
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .						

Tabelle 8.11: ANOVA Zwischenmodell I mit imputierten fehlenden Werten

	McF.R^2	CSR^2	Nag.R^2	AIC	BIC	κ
RM	0.58	0.39	0.68	1'202	1'979	*Inf*
ZI ohne	0.34	0.29	0.46	1'040	1'062	$\approx$ 100'000
ZI mit	0.25	0.19	0.33	1'371	1'389	$\approx$ 285'000

RM: Referenzmodell
ZI ohne: Zwischenmodell I ohne Imputation
ZI mit: Zwischenmodell I mit Imputation

Tabelle 8.12: Gütekriterien des Zwischenmodells I mit imputierten fehlenden Werten

8.3.2 Zwischenmodell II: Variablen des Baums

Abgesehen von der unabhängigen Modellierung der Teilnahmewahrscheinlichkeit wurden die Variablen des Baums in einem Probit-Modell verwendet, um die Teilnahmewahrscheinlichkeit zu modellieren und zu prognostizieren. Das Modell ist unabhängig von den Informationskriterien nicht brauchbar, da es zu einer perfekten Vorhersage kommt. Im GLM wurden die Wechselwirkungen nicht berücksichtigt. Tabelle 8.14 listet die Informationskriterien auf.

Das sehr hohe κ zeigt, dass die Multikollinearitäten sehr hoch sind. Den Pseudo-Bestimmtheitsmassen ist daher nicht zu trauen. Verzichtet man im Modell auf die Befragten, die relativ viele fehlende Werte aufweisen, bricht

		Referenzmodell		ohne Imputation		mit Imputation	
	korrekt	kT	T	kT	T	kT	T
kT	1'752	1'672	80	1'671	81	1'741	11
T	316	114	202	174	142	281	35

Tabelle 8.13: Prognose der Teilnahme im Zwischenmodell

McFadden R^2	0.50
Cox-Snell R^2	0.35
Nagelkerke R^2	0.35
AIC	995
BIC	1'109
κ	$\approx 7.5 * 10^6$

Tabelle 8.14: Informationskriterien des Zwischenmodells II

das Modell zusammen und weisst nur noch je nach verwendeten Variablen Pseudo-Bestimmtheitsmasse zwischen 0.01 und maximal 0.09 auf.

Tabelle 8.15: Vergleichsmodell

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	-3.5507	1086.2301	-0.00	0.9974	
N904051	2.3143	0.5259	4.40	0.0000	***
N904052	0.7089	0.4906	1.44	0.1485	
N904053	-0.1232	0.5763	-0.21	0.8308	
N904054	0.6968	0.6323	1.10	0.2704	
N904055	-18.4066	3298.2979	-0.01	0.9955	
N904056	-18.9717	9440.8112	-0.00	0.9984	
N904057	-11.1619	17730.3699	-0.00	0.9995	
N904058	-18.5107	17730.3699	-0.00	0.9992	
N9040599	6.2761	687.9479	0.01	0.9927	
f911001	-4.1711	1.7238	-2.42	0.0155	*

Fortsetzung nächste Seite . . .

Tabelle 8.15: Vergleichsmodell

	Estimate	Std. Error	t value	Pr(> |t|)	
f911002	-3.5489	1.6093	-2.21	0.0274	*
f911003	-3.5972	1.6051	-2.24	0.0250	*
f911004	-3.4759	1.6194	-2.15	0.0318	*
f911005	-3.6759	1.6197	-2.27	0.0232	*
f911006	-3.4350	1.6152	-2.13	0.0334	*
f9110099	-2.5476	1.8667	-1.36	0.1723	
N00105	0.0340	0.0092	3.69	0.0002	***
f851003	3.7804	1086.2283	0.00	0.9972	
f851004	5.0666	1086.2282	0.00	0.9963	
f851005	4.4526	1086.2282	0.00	0.9967	
f851006	4.9031	1086.2282	0.00	0.9964	
f8510099	-0.7293	9469.4301	-0.00	0.9999	
f913002	-0.4136	0.2390	-1.73	0.0835	.
f913003	0.3573	0.2707	1.32	0.1869	
f913004	0.2006	0.3803	0.53	0.5978	
f913005	-0.2809	0.4267	-0.66	0.5103	
f913006	0.2088	0.6493	0.32	0.7478	
f9130099	2.1836	0.4675	4.67	0.0000	***
f903002	0.3220	0.2539	1.27	0.2047	
f903003	3.2498	0.9278	3.50	0.0005	***
f903004	-0.2093	0.3308	-0.63	0.5269	
N85200	0.0144	0.0056	2.58	0.0097	**
f001102	0.0848	0.1946	0.44	0.6631	
f034052	0.1318	0.5545	0.24	0.8121	
f034053	0.5357	0.5307	1.01	0.3128	
f034054	0.4450	0.5355	0.83	0.4060	
f034055	-0.1349	0.5707	-0.24	0.8131	
f0340599	-14.8860	6047.7115	-0.00	0.9980	
f034092	-0.5093	0.5995	-0.85	0.3956	
f034093	-0.8897	0.5722	-1.56	0.1199	
f034094	-0.7677	0.5783	-1.33	0.1843	

Fortsetzung nächste Seite . . .

Tabelle 8.15: Vergleichsmodell

	Estimate	Std. Error	t value	Pr(> |t|)	
f034095	-1.1199	0.6058	-1.85	0.0645	.
f0340999	0.7117	1.4991	0.47	0.6350	
f033072	0.0386	0.2094	0.18	0.8536	
f033073	0.1966	0.2544	0.77	0.4398	
f033074	-0.0539	0.3203	-0.17	0.8664	
f033075	-0.5530	0.4801	-1.15	0.2494	
f0330799	-17.5540	4016.3500	-0.00	0.9965	
f035022	0.9336	0.4244	2.20	0.0278	*
f035023	0.3673	0.3995	0.92	0.3579	
f035024	0.5597	0.4100	1.37	0.1722	
f035025	-0.2545	0.4774	-0.53	0.5939	
f0350299	112.0045	7259.8850	0.02	0.9877	
noNA	-1.8511	0.1780	-10.40	0.0000	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .

Nur relativ wenige Faktoren haben einen signifikanten Einfluss im Probit-Modell. Ein differenzierteres Bild liefert die ANOVA, siehe Tabelle 8.16.

Tabelle 8.17 zeigt, dass der mit 174 (8.3 %) etwas weniger Befragte falsch klassifiziert als das Probit-Modell mit 211 (11.1 %) Falsch-Klassifizierungen. Der Unterschied mag nicht sehr gross erscheinen, wird aber deutlicher, wenn man umformuliert in der Baum «erkennt» 66.8 % der 316 Panel-Teilnehmer richtig, die Regression dagegen nur 55.1 %. Es wird daher empfohlen, zur Bestimmung auch der *propensity scores* eher den Baum zu verwenden, als eine Probit-Regression.

In Abbildung 8.3 sind die *propensity scores* als prognostizierte Teilnahmewahrscheinlichkeiten aus dem Baum und aus dem Probit-Model gegeneinander geplottet. Befragte, die tatsächlich am Panel teilnehmen sind durch rote Kreise dargestellt, diejenigen, die nicht teilnehmen durch blaue.

Vergleicht man die resultierenden Werte zwischen Baum und Probit-Model, zeigt sich, dass es einige Beobachtungen gibt, für die der Baum und das Probit-Modell völlig unterschiedliche *propensity scores* bestimmen.

	Df	Deviance	Resid. Df	Resid. Dev	P(> |χ|)	
N90405	9	261.53	2058	1506.79	0.0000	***
f91100	7	33.47	2051	1473.32	0.0000	***
N00105	1	3.68	2050	1469.65	0.0552	.
f85100	5	119.78	2045	1349.87	0.0000	***
f91300	6	23.72	2039	1326.14	0.0006	***
f90300	3	6.28	2036	1319.86	0.0987	.
N85200	1	4.29	2035	1315.57	0.0383	*
f00110	1	0.09	2034	1315.48	0.7687	
f03405	5	12.60	2029	1302.88	0.0274	*
f03409	5	2.60	2024	1300.28	0.7611	
f03307	5	5.83	2019	1294.45	0.3229	
f03502	5	26.33	2014	1268.12	0.0001	***
noNA	1	384.58	2013	883.54	0.0000	***

Signif. Niveau: 0 *** 0.001 ** 0.01 * 0.05 .

Tabelle 8.16: ANOVA Vergleichsmodell

		Baum		Zwischenmodell	
	korrekt	kT	T	kT	T
kT	1'752	1'685	67	1'664	88
T	316	105	211	142	174

Tabelle 8.17: Prognose der Teilnahme des Zwischenmodells

Benutzt man das beschriebene Zwischenmodell als Ausgangsmodell für eine weitere Modellreduktion, bleiben folgende Variablen in dem resultierenden Modell übrig: `N90405`, `f91100`, `f85100`, `f91300`, `f03502` und `f85100`. Es resultieren Informationskritierien gemäss Tabelle 8.18.

Die Pseudo-Bestimmtheitsmasse sind also gut, obwohl die Multikollinearitäten sehr hoch sind. Das eigentlich entscheidende Kriterium ist allerdings, wie gut die Prognose der Teilnahme tatsächlich ist.

Nur rund die Hälfte all derer, die an der Befragung teilnehmen werden also vom reduzierten Zwischenmodell richtig prognostiziert, Tabelle 8.19. Es

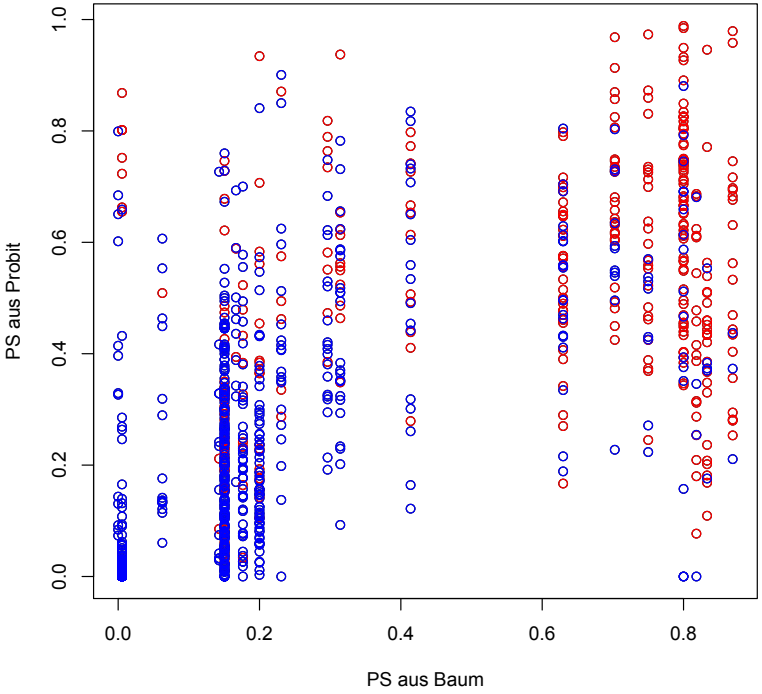


Abbildung 8.3: Vergleich der *propensity scores* aus dem Baum und aus dem Probit Modell

ist damit untauglich.

	McF.R^2	CSR^2	Nag.R^2	AIC	BIC/SBC	κ
Referenzmodell	0.30	0.34	0.45	1'800	1'923	3'798
Zwischenmodell	0.12	0.13	0.19	993	1'081	4'714
Reduziertes Zwischenmodell	0.46	0.33	0.57	1'021	1'073	≈2.5 Mio

Tabelle 8.18: Gütekriterien des reduzierten Zwischenmodells

		Baum		Zwischen-modell		red. Zwischen-modell	
	korrekt	kT	T	kT	T	kT	T
kT	1'752	1'685	67	1'664	88	1'653	99
T	316	105	211	142	174	157	159

Tabelle 8.19: Prognose der Teilnahme des reduzierten Zwischenmodells

8.4 Fazit: Entscheidung für das Baummodell

Zwei Gründe sind ausschlaggebend dafür, sich für den Baum als Methode zu entscheiden, die Teilnahmewahrscheinlichkeit vorherzusagen und damit die *propensity scores* zu bestimmen. Zum einen ist die Erklärungskraft eines Baumes deutlich höher, was insbesondere dann gilt, wenn die Anzahl der zur Verfügung stehenden erklärenden Variablen beschränkt ist. Zum zweiten haben die vorgeschlagenen Variablen innerhalb der Probit-Modelle, keine hohe Erklärungskraft. Eine grosse Ausnahme bildet die Information «Anzahl fehlender Werte», die allerdings zu einer perfekten Vorhersage bei den probit-Modellen führt. Wird auf diese Information verzichtet und z. B. fehlende Werte imputiert, sinkt die Erklärungskraft des Modells deutlich.

9 Gewichtung mittels Teilnahmewahrscheinlichkeiten aus Baum

► Nachdem in den beiden vorangegangenen Kapitel gezeigt wurde, wie die Teilnahmewahrscheinlichkeit modelliert werden konnte, sollen in diesem Kapitel Gewichte abgeleitet und angewendet werden. Zunächst geschieht dies mit dem vollständigen, in einem weiteren Abschnitt mit einem reduzierten Datensatz. Um das PSA besser zu verstehen wird im sich anschliessenden Abschnitt die die Teilnahmewahrscheinlichkeit modifiziert, um so erste kleine Simulationen zu erlauben. Im letzten Abschnitt dieses Kapitels werden basierend auf den modellierten Teilnahmewahrscheinlichkeiten der vergangenen Kapitel die R-Indikatoren bestimmt und diskutiert. ◄

9.1 PSA mit vollständigem Datensatz

Mit Hilfe des Baums kann die Teilnahmewahrscheinlichkeit geschätzt werden, siehe Kapitel 7 ab S. 119. Wie aus der Struktur des Baums folgt, sind die berechneten Wahrscheinlichkeiten nicht stetig verteilt, sondern haben höchstens so viele Ausprägungen wie Blätter. In diesem Fall konnten 19 verschiedene Wahrscheinlichkeiten, d h. Propensity Scores bestimmt werden.

2 Gewichtungsarten möglich

Diese Propensity Scores können auf zweierlei Art als Gewicht herangezogen werden. Zum einen kann unmittelbar das Reziprok der Teilnahmewahrscheinlichkeit ($1/\rho$) als Gewicht verwendet werden. Dies ist allerdings bei einem Baum Modell schwierig, da Teilnahmewahrscheinlichkeiten von Null geschätzt werden können. Hier ist dies bei 13 Beobachtungen der Fall, die allerdings auch alle tatsächlich nicht teilnehmen. Nicht definierte Gewichte von $1/0$ kommen also nicht vor.

Eine andere Möglichkeit, Gewichte zu konstruieren, besteht darin, die Teilnahmewahrscheinlichkeiten weiter zu klassifizieren. Das bedeutet, dass

Befragte mit einer ähnlichen Teilnahmewahrscheinlichkeit in Gruppen zusammengefasst werden. Die Gewichte leiten sich aus dem Vergleich der Gruppengrösse zwischen Teilnehmenden und der Grundgesamtheit ab; siehe dazu auch Kapitel 3, ab S. 25.

Die beiden Möglichkeiten sollen folgend «direkt» und «klassifiziert» genannt werden. Die direkte Methode bietet den Vorteil, dass sie gewöhnlich besser ist, da die Gewichte differenzierter sind. Der Nachteil liegt darin, dass die Methode sehr ausreisseranfällig sein kann. Befragte mit einer sehr geringen Teilnahmewahrscheinlichkeit erhalten sehr hohe Gewichte.

Der Anteil derjenigen mit sehr kleiner Teilnahmewahrscheinlichkeit ist in der Grundgesamtheit aller CATI-Befragten gross (58 % haben eine Wahrscheinlichkeit von weniger als einem Prozent), bei den Panelisten ist deren Anteil dagegen sehr klein (7 Probanden, d. h. 2.2 %). Diese sieben Beobachtungen erhalten daher sehr hohe Gewichte. Das Verhältnis des höchstem zum niedrigsten Gewicht entspräche rund 131. Die Summe der Gewichte aller 316 Beobachtungen im Panel beträgt 1'834. Die Summe der Gewichte der sieben Beobachtungen im Panel mit den kleinsten Teilnahmewahrscheinlichkeiten entspricht 1'060. Sie repräsentieren damit 58 % des Gewichts der gesamten Stichprobe.

Zum Vergleich der Verteilung der Teilnahmewahrscheinlichkeiten und Gewichte zwischen Grundgesamtheit und Panelisten siehe auch Tabelle 9.1.

Um dieses Problem zu umgehen, wurden die Gewichte beim zweitkleinsten Gewicht (16) trunkiert. Das Verhältnis vom grössten zum kleinsten Gewicht reduziert sich dann auf akzeptable 13.9. Im Panel verbleiben so 17 Gruppen mit unterschiedlichen Gewichten. Welches Gewichts-Verhältnis akzeptabel ist, ist nicht allgemeingültig definierbar. Dass das Verhältnis sehr gross werden kann, ist eigentlich plausibel und möglicherweise realistisch. Allerdings ist es nicht wünschenswert, wenn 2.2 % der Befragten 58 % eines Anteils bestimmen.

Klassifiziert in 5 Gruppen

Es wurden gemäss dem Vorschlag von Cochran (1968) fünf Gruppen mit ähnlichen Teilnahmewahrscheinlichkeiten gebildet. Diese Klassifizierung wurde unterschiedlich vorgenommen. Jeweils gemeinsam haben alle Klassifikationen, dass ausgehend von den Panelisten möglichst ähnlich grosse Gruppen gebildet wurden.

Eine Möglichkeit ist es, Gruppen so zu bilden, dass all diejenigen mit gleicher Teilnahmewahrscheinlichkeit auch sicher in derselben Gruppe sind.

Tabelle 9.1: Gewichte gemäss direkter Methode

		Häufigkeit	
ρ	Gewicht	GG	S
0.870	1.15	23	20
0.833	1.20	30	25
0.812	1.22	22	18
0.800	1.25	80	64
0.752	1.33	32	24
0.704	1.42	37	26
0.629	1.59	54	34
0.413	2.42	29	12
0.314	3.18	35	11
0.297	3.37	27	8
0.230	4.33	26	6
0.200	5.00	80	16
0.176	5.67	51	9
0.167	6.00	12	2
0.150	6.66	206	31
0.143	7.00	14	2
0.062	16.00	16	1
0.007	151.43	1'073	7

Diese Variante ist in Tabelle 9.2 dokumentiert und zeigt die Verteilung der Gruppengrösse mit den jeweiligen Gewichten.

Das Verhältnis zwischen grösstem und kleinstem Gewicht beträgt jetzt 17.9. Ein Nachteil dieser Variante besteht darin, dass es insbesondere bei der Modellierung der Teilnahmewahrscheinlichkeit mit einem Baum zu einzelnen sehr grossen oder sehr kleinen Gruppen kommen kann. In diesem Fall hier ist das Verhältnis zwischen grösster und kleinster Gruppe nicht sehr extrem, bei Versuchen die Wirkung der Gewichtungen mit alternativen Baum-Modellen nachzuvollziehen, gab es gelegentlich sehr extreme Verteilungen.

Eine Alternative zu diesem Vorgehen besteht darin, dass um die unterschiedlichen Gruppengrössen und damit Gewichte auszugleichen, zugelassen wurde, dass Befragte mit gleich grossen Teilnahmewahrscheinlichkeiten

Gruppe	Grösse in GG	S	gerundetes indiv.Gewicht
1	1'425	68	3.65
2	171	71	0.41
3	69	50	0.24
4	80	64	0.22
5	75	63	0.20

Tabelle 9.2: Vergleich der Gewichtungen

auf unterschiedliche Gruppen verteilt wurden, mit dem Zeil, dass die Gruppengrösse bei den Panelisten immer 1/5∗Anzahl Panelisten beträgt. Dieses Vorgehen ist nicht nur beliebig, sondern hat sich auch noch weniger bewährt, als die eben beschriebene Methode.

Tabelle 9.3 listet für die ursprünglich implementierten Kontrollvariablen «Präferenz für einen Supermarkt» und «Besitz eines SBB-Abonnements» die Verteilung der Anteile auf der Ebene der Webnutzer auf[1] und vergleicht diesen mit den geschätzten Werten aus der Stichprobe, wobei die Anteile jeweils ungewichtet und direkt bzw. klassifiziert gewichtet aufgeführt sind.

Alle gewichteten Anteile liegen innerhalb des Konfidenzintervalle der ungewichteten Anteile. Es besteht bei keinem Anteil ein signifikanter Unterschied zwischen den gewichteten und dem ungewichteten Wert. Eine Ausnahme bildet der Besitz einer Mehrfahrtenkarte, wobei der ungewichtete Anteil signifikant über dem tatsächlichen Anteil liegt. Beide Korrekturversionen können das aber wieder ausgleichen.

Die vorgeschlagen Kontrollvariablen unterscheiden sich nicht stark zwischen Panelisten und nicht-Panelisten. Der Responsemechanismus hat also keinen Einfluss auf die interessierende Variable, es handelt sich bezüglich der Kontrollvariablen um einen missing-completely-at-random Ausfallprozess.

Zur nochmaligen Überprüfung wurden die Kontrollvariablen in einem Linearen Probit-Modell mit Hilfe der Variablen, die zur Schätzung der Teil-

1 Das heisst auf der Ebene all derjenigen, die mindestens ein Mal pro Woche das Internet benutzen und nicht in der Marktforschung arbeiten.

	Tatsächlich Ebene Web	gewichtet		
		ungewichtet	direkt	klassifiziert
Migros	64.4	64.7	64.4	63.9
Coop	55.6	54.0	52.7	54.8
Denner	7.5	7.3	7.1	8.0
Andere	20.2	18.4	15.5	19.2
Halbtax	38.6	42.4	40.2	37.6
GA	12.9	12.6	11.5	10.6
Verbund-Abo	11.1	14.2	14.4	15.6
Strecken-Abo	7.9	6.1	6.3	6.6
Mehrfahrtenkarte	11.6	17.1	11.9	9.4
Gleis 7	2.1	1.9	3.5	5.2
Andere	2.6	2.9	2.9	3.0
Keine	37.1	32.7	33.7	35.4

Tabelle 9.3: Kontrollvariablen I

nahmewahrscheinlichkeiten mit Baum herangezogen wurden, geschätzt.

Bei den Kontrollvariablen handelt es sich ja um eine Reihe dichotomer Varaiblen («Ich präferiere Supermarkt X oder nicht» bzw. «Ich bestitze das Bahn-Abo Y oder nicht»). Daher waren zur Schätzung aller einzelner Kontrollvariablen Probit-Modelle angemessen. Die Güte der Probitmodelle sind ein Indikator dafür, wie gut die Kovariate des Baums die jeweilige Variable erklären, d. h. auch wie stark die Teilnahmewahrscheinlichkeit mit der jeweiligen Variable korreliert. Dies in Form einer Regression zu machen, erlaubt auch den Einfluss der einzelnen Kovariate bei Bedarf jeweils untersuchen zu können.

Tabelle 9.4 listet die beiden wichtigen Informationskriterien zur jeweiligen Modellgüte auf.

Keine der Variablen, die zur Kontrolle in die Befragung eingefügt wurden, lassen sich gut mit Hilfe der Variablen des Baums erklären. Das Pseudo-Bestimmtheitsmass Nagelkerke R^2 ist bei bei allen Variablen bis auf `f03105` niedrig. Das ist auch die einzige Variable, bei der eine signifikante Verbesserung durch die Gewichtung erzielt werden konnte. κ als Mass der Multikolinerität ist bei allen Modellen sehr hoch. Es werden andere als die

Variable		Nagelkerke R^2	$\kappa > 100$
f03001	Migros	0.05	TRUE
f03002	Coop	0.05	TRUE
f03003	Denner	0.08	TRUE
f03004	Andere	0.05	TRUE
f03101	Halbtax	0.09	TRUE
f03102	GA	0.12	TRUE
f03103	Verbund-Abo	0.08	TRUE
f03104	Strecken-Abo	0.14	TRUE
f03105	Mehrfahrtenkarte	0.38	TRUE
f03106	Gleis 7	0.14	TRUE
f03107	Andere	0.15	TRUE
f03108	keine	0.08	TRUE

Tabelle 9.4: Prognose der Kontrollvariablen I

ursprünglich geplanten Variablen als Kontrollvariablen benötigt.

Wie in Tabelle 9.5 aufgelistet gibt es allerdings Variablen, die sich signifikant zwischen Panelisten und nicht-Panelisten unterscheiden. Die Tabelle zeigt wiederum die tatsächliche Verteilung der Grundgesamtheit (also der CATI-Befragung) und die gewichteten Schätzungen. (Die Anzahl fehlender Werte wurde nicht als eigene Ausprägung der Variablen behandelt!)

Tabelle 9.5: Kontrollvariablen II

f85100	Internetnutzung privat				
Levels	3	4	5	6	
CATI	8.77	23.44	28.37	39.42	
Stichprobe	2.85	25.32	22.15	49.68	
GewReziprok	2.95	27.65	24.03	45.37	
GewKlassifiziert	1.75	29.36	26.63	42.25	
f03505	Wichtigkeit Regeln				
Levels	1	2	3	4	5
CATI	4.24	10.78	31.57	32.51	20.9

Tabelle 9.5: Kontrollvariablen II

Stichprobe	1.94	5.5	28.48	43.37	20.71	
GewReziprok	1.58	4.99	28.02	44.29	21.11	
GewKlassifiziert	1.69	4.98	28.03	42.81	22.51	
f03404	Wichtigkeit Abwechslung					
Levels	1	2	3	4	5	
CATI	0.99	4.58	22.74	39.02	32.67	
Stichprobe	0.65	6.47	24.27	37.22	31.39	
GewReziprok	0.35	5.02	20.2	35.38	39.04	
GewKlassifiziert	0.21	4.01	17.43	32.08	46.27	
f03501	Daten vertraulich bei MaFo					
Levels	1	2	3	4	5	
CATI	4.29	10.51	30.25	27.45	27.5	
Stichprobe	2.27	8.09	27.18	31.72	30.74	
GewReziprok	2.05	7.46	28.03	27.93	34.52	
GewKlassifiziert	1.9	5.75	27.6	26.3	38.45	
f03502	Umfragen sind abwechslungsreich					
Levels	1	2	3	4	5	
CATI	12.21	19.31	35.59	23.6	9.3	
Stichprobe	3.88	22.65	38.51	27.83	7.12	
GewReziprok	4.72	15.8	40.12	31.29	8.07	
GewKlassifiziert	5.81	13.65	39.25	33	8.29	
f91200	Haushaltseinkommen					
Levels	1	2	3	4	5	6
CATI	5.96	19.19	26.58	19.67	18.95	9.65
Stichprobe	3.54	15.49	29.65	19.47	23.45	8.41
GewReziprok	4.12	15.34	31.86	22.63	18.68	7.38
GewKlassifiziert	5.55	16.56	33.4	18.64	17.98	7.88

Die Gewichtungen führen nicht zu einer Verbesserung der Schätzer! Unabhängig davon, inwieweit die Unterschiede tatsächlich signifikant sind, zeigt sich, dass die ungewichtete Schätzung der Verteilungen aus der Stichprobe häufig sogar näher am echten Wert der CATI liegt, als die gewichteten

Schätzungen.

Jetzt zeigt sich doch, dass die Modellierung der Teilnahmewahrscheinlichkeit nicht gut gewesen ist. Jedenfalls lässt sich der Bias, der aus unterschiedlichem Antwortverhalten von Panelisten und nicht-Panelisten resultiert, nicht ohne weiteres korrigieren.

Die wichtigste Ursache ist sicherlich, dass die Unterschiede zwischen Teilnehmern am Panel und den CATI-Befragten empirisch nicht sehr gross sind. Eine Gewichtung verbietet sich schon aus diesem Grunde.

9.2 PSA mit verkleinertem Datensatz GG^{816}

Um zu überprüfen, ob die Methode im Kontext des Web-Panels doch zu einer brauchbaren Reduktion des Bias führen kann, wurde als neue Grundgesamtheit alle Befragten der CATI-Befragung bestimmt, die nicht im Boost-Sample sind, da bei diesen Befragten sehr viel weniger fehlende Werte vorhanden sind. Dabei handelt es sich um 816 Befragte[2] (GG^{816}).

Aus GG^{816} haben 138 an der Panel-Befragung teilgenommen. Die Teilstichprobe soll kurz S^{basis} genannt werden. Mit Hilfe der Daten dieser Befragten können verschiedene ähnlich gute Modelle entwickelt werden. Exemplarisch soll eins vorgestellt werden, dessen Variablen auch in den Überprüfungen der Methode in den folgenden Abschnitten verwendet werden sollen.

Auch wenn das Vorgehen in der Praxis ungeeignet ist, wurden zunächst alle Beschränkungen bezüglich der Auswahl der Variablen aufegegeben, auch um die Methode besser zu verstehen.

9.2.1 Teilnahmewahrscheinlichkeit schätzen

Neue Modellierung

Um die Teilnahmewahrscheinlichkeit zu schätzen, wurde völlig neu modelliert. Dies war notwändig, da es sich beim verkleinerten Datensatz bezüglich der Schätzung der Teilnahmewahrscheinlichkeiten um einen völlig unterschiedlichen Datensatz handelt. Da die fehlenden Werte keine wichtige Rolle mehr spielen, konnte die Teilnahmewahrscheinlichkeit erfolgreich mittels Probit-Modellen geschätzt werden. Diese beiten im Vergleich zu Bäumen nicht nur den Vorteil, dass die Indikatoren zur Beurteilung der Güte ausgereifter sind (siehe Liste in Abschnitt 8, S. 137).

In dem neu gefundenen Modell dominieren Variablen aus der Gruppe der Rational-Choice Fragen (`f035xx` und `f85100`, vergleiche auch Tabelle 9.6). `f03308` ist die einzig verbleibende Variable aus dem Big-Five-Inventar und `f90300` die einzige demografische Variable. Letztere musste jedoch dichotomisiert werden (verheiratet / sonstiger Status). Weitere demografische Variablen wie Alter und Geschlecht konnten keinen zusätzlichen Informationsgewinn im Modell beitragen und wurden daher nicht aufgenommen.

2 Tatsächlich bleiben nach Abzug der im Boost Befragten 830 Probanden übrig. Allerdings haben 14 von ihnen sehr viele fehlende Werte, weshalb sie nicht berücksichtigt wurden.

Tabelle 9.6: Modellierung der Teilnahmewahrscheinlichkeit mit S^{basis}

	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	-6.4146	0.9782	-6.56	0.0000	***
f035042	0.4543	0.4725	0.96	0.3364	
f035043	0.8976	0.4462	2.01	0.0443	*
f035044	1.0513	0.4425	2.38	0.0175	*
f035045	1.4633	0.4933	2.97	0.0030	**
f035022	1.0238	0.5249	1.95	0.0511	.
f035023	1.5589	0.4949	3.15	0.0016	**
f035024	1.5950	0.5068	3.15	0.0016	**
f035025	0.7447	0.5971	1.25	0.2124	
f851004	1.7621	0.7668	2.30	0.0216	*
f851005	2.0442	0.7537	2.71	0.0067	**
f851006	2.4407	0.7430	3.29	0.0010	**
f03308r2	0.0918	0.2525	0.36	0.7162	
f03308r3	0.8982	0.2725	3.30	0.0010	***
f03308r4	0.1776	0.4084	0.43	0.6638	
f90300r2	0.6252	0.2007	3.11	0.0018	**

Signif. Niveau: 0 *** 0.001 ** 0.01 * 0.05 .

Geringe Erkläungskraft

Die Erklärungskraft des Modells ist nicht sehr hoch, aber auch nicht schlecht. Die Pseudo-Bestimmtheitsmasse sind alle relativ niedrig, was positiverweise auch für die Multikollinearitäten gilt, siehe Tabelle 9.7.

McF.R^2	CSR^2	Nag.R^2	AIC	BIC/SBC	κ
0.11	0.09	0.15	693	706	21

Tabelle 9.7: Gütekriterien des Modells mit S^{basis}

Dass sich die Variablen dieses Modells doch deutlich von denjenigen unterscheiden, die im Baum verwendet wurden, ist bezüglich der Interpretation überraschend. Offensichtlich wurden noch nicht die wichtigen Variablen gefunden, die das Teilnahmeverhalten unter verschiedensten Bedingungen

gut prognostizieren können. Eine weitere Suche auch jenseits von Scheatz-Skala, Big-Five-Inventory, Rational-Choice und demografischen Fragen wird notwändig sein, um die Methode erfolgreich anwenden zu können.

Vielleicht sind allerdings auch die Rational-Choice Fragen gar nicht so schlecht, sondern nur die Datenlage etwas ungeschickt, da die Unterschiede bei den Kontrollvairablen, wie jetzt schon häufig geschrieben, zu gering sind.

9.2.2 Resultierende Teilnahmewahrscheinlichkeiten und Gewichtungen

Die kleinste Teilnahmewahrscheinlichkeit beträgt 0.18 %, die grösste 54.88 %, die durchschnittliche Teilnahmewahrscheinlichkeit beträgt 15.00 %. Da die kleinste Teilnahmewahrscheinlichkeit all derer, die tatsächlich an der Befragung teilgenommen haben, 3.77 % beträgt (bei einem Maximum von 53.98 %) ist das Verhältnis von grösstem (26.56) zum kleinstem Gewicht (1.85) bei direkter Gewichtungsmethode 14.3. Dieser Betrag ist nicht problematisch. Die Verteilung ist nochmals in Tabelle 9.8 abgebildet.

Minimum	0.0018
1. Quartil	0.0793
Median	0.1503
Arith. Mittel	0.1697
3. Quartil	0.2320
Maximum	0.5488

Tabelle 9.8: Verteilung der Teilnahmewahrscheinlichkeiten des Modells mit S^{basis}

Die Verteilung entspricht dem Histogramm in Abbildung 9.1:

Als Vergleich wurden die Gewichte auch wieder via Klassifzierung der Teilnahmewahrscheinlichkeiten ausprobiert. Bei dieser Methode mit fünf Klassen ergibt sich ein Verhältnis aus grösstem zu kleinstem Gewicht von 5.71.

Wendet man die Gewichte auf die (vereinfachte) Supermarktpräferenz[3] an, ergeben sich Schätzungen der Anteile gemäss Tabelle 9.18.

Kontrollvariablen

3 Im Fragebogen handelt es sich um eine Frage mit Mehrfachantwortmöglichkeit. Um die folgenden Ausführungen zu vereinfachen, wurde so umkodiert, dass jeder Befragter genau einem Supermarkt zugeordnet wird. Im Falle mehrerer Angaben gilt folgende Hierarchie: Migros, Coop, Denner, Sonstige.

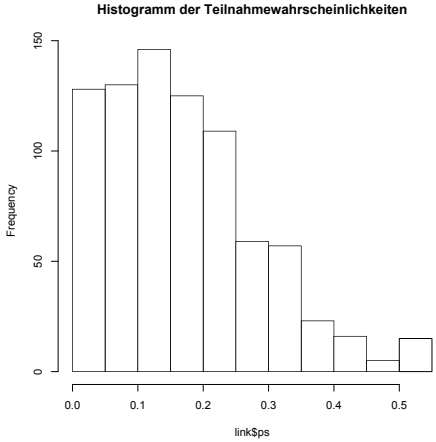


Abbildung 9.1: Histogramm der Teilnahmewahrscheinlichkeiten im Modell mit S^{basis}

Gruppe	Grösse in GG^{816}	Grösse in S^{basis}	gerundetes Gewicht
1	67	27	0.42
2	71	28	0.43
3	127	28	0.77
4	166	28	1.00
5	382	27	2.40

Tabelle 9.9: Klassifizierte Gewichte in S^{basis}

Keine Biasreduktion

Abgesehen von dem Umstand, dass keiner der Unterschiede signifikant ist, führen beide Arten der Gewichtung nicht zu einer Verbesserung der Anteilsschätzer im Vergleich zur einfachen ungewichteten Stichprobe S^{basis}. Im Gegenteil, die Tendenzen des Bias scheinen sich noch zu verstärken.

Das kann mindestens durch zwei Umstände bedingt sein. Zunächst kann es einfach Zufall sein, dass es überhaupt eine Abweichung gibt. Wie in Abbildung 8.3 zu sehen ist, gibt es keine starke Korrelation zwischen der

	Anteil ungewichtet		Anteil gewichtet	
	in GG^{816}	in S^{basis}	direkt	kategorisiert
Migros	64.1	60.1	56.2	57.3
Coop	26.7	29.0	29.3	29.8
Denner	1.5	2.9	4.7	4.5
Sonstige	7.6	8.0	9.8	8.4

Tabelle 9.10: Vergleich der Gewichtungsarten Modell S^{basis}

Supermarktpräferenz und der Teilnahmewahrscheinlichkeit. Mithin ist es möglich, dass die Abweichung des gewichteten Schätzers nur zufällig grösser ist, als der des ungewichteten Schätzers. Die Konfidenzintervall der ungewichteten Schätzer (a =95 %, Endlichkeitskorrektur berücksichtigt)[4] können Tabelle 9.11 entnommen werden und zeigen, dass sie doch recht gross sind.

		Grenze	
	Anteil	Unten	Oben
Migros	56.2	48.7	63.7
Coop	29.3	22.4	36.2
Denner	4.7	1.5	7.9
Sonstige	9.8	5.3	14.3
Halbtax	18.1	12.2	24.0
Generalabonnement	10.1	5.5	14.7
Verbund-Abonnement	12.3	7.3	17.3
Strecken-Abonnement	4.3	1.2	7.4
Mehrfahrtenkarte	17.9	12.1	23.7
Gleis-7 / Anderes	5.0	1.7	8.3
Keines	30.4	23.4	37.4

Tabelle 9.11: Konfidenzintervalle der ungewichteten Schätzer in S^{basis}

4 Das Konfidenzintervall wurde berechnet als

$$a - Konf(\pi) = \left[\hat{p} \pm z_{\frac{1+a}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}(1 - \frac{n}{N}}\right]_a$$

Ausnahmslos alle gewichteten Schätzer liegen klar innerhalb der Konfidenzintervalle. Das gilt auch für die zweite Kontrollvariable, dem Besitz eines SBB-Abonements.

Ein weiterer Grund dafür, dass die gewichteten Schätzer nicht zu einer Verbesserung führen, kann sein, dass die Informationen, die im Modell zur Prognose der Teilnahmewahrscheinlichkeiten verwendet werden, nicht ausreichend sind. Das ist nicht zuletzt an den niedrigen Pseudo-Bestimmtheitsmassen zu erkennen. Ausserdem ist der Unterschied zwischen der Grundgesamtheit GG^{816} und den Panelisten S^{basis} weiterhin sehr klein. Das gilt nicht nur für das Beispiel der Supermarktpräferenz als abhängiger Variable, sondern auch für andere, hier nicht dokumentierte Variablen.

Erstaunlicherweise führt die Verwendung von Gewichten bei der zweiten Kontrollvariable `sbb`[5] zu einer leichten Verbesserung der Schätzer für die Anteile!

	Anteil in GG^{816}	in S^{alt}	in S^{plus}
Halbtax	18.9	20.3	18.1
Generalabonnement	10.9	10.1	11.0
Verbund-Abonnement	8.6	12.3	10.6
Strecken-Abonnement	7.4	4.3	4.4
Mehrfahrtenkarte	12.3	17.9	16.8
Gleis-7 / Anderes	4.9	5.0	6.4
Keines	36.9	30.4	32.8

Tabelle 9.12: Verteilung der SBB-Abonnements

Man kann zwar vermuten, dass die Tendenz im Unterschied möglicherweise daher kommt, dass `sbb` höher mit der Teilnahmewahrscheinlichkeit korreliert, als die Supermarktpräferenz, aber das trifft nicht zu, wie man in Abbildung 8.3 sehen kann. Vermutlich hat die Gewichtung gar nicht zu einer «echten» Verbesserung geführt, sondern nur zu einem Schätzer, der dem

5 Die Variable `sbb` wurde so gebildet, dass mögliche Mehrfachantworten nicht mehr erfasst wurden. Dies ist aus Gründen der rechnerischen Vereinfachung geschehen. Die neu gebildete Variable umfasst nur sieben statt acht Ausprägungen, da Gleis-7 uns sonstige Abonnements so selten waren, dass die Kategorien zusammengefasst wurden. Die Hierarchie der Zuordnung bei Mehrfachnennung entspricht der Nennung in Tabelle 9.12 in umgekehrter Reihenfolge.

ungewichteten sehr ähnlich ist. Beide unterscheiden sich nicht signifikant.

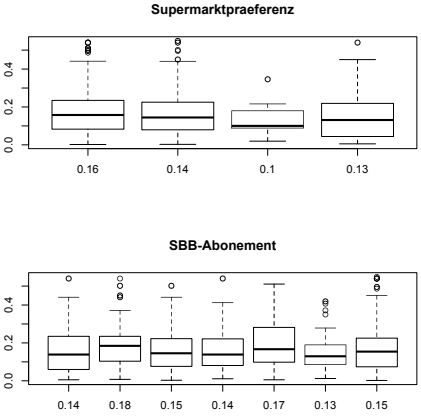


Abbildung 9.2: Vergleich der *propensity scores* zwischen Supermarktpräferenz und SBB-Abonnement

9.3 PSA mit modifizierter Teilnahmewahrscheinlichkeit

9.3.1 Das einfache Modell

Damit es tatsächlich einen Unterschied zwischen Der Grundgesamtheit und den Panelisten gibt, wurde die Teilnahmewahrscheinlichkeit manipuliert.

Neben den 141 Befragten, die tatsächlich bereit gewesen sind, am Panel teilzunehmen (S^{alt}), wurden zusätzliche Befragte zufällig ausgewählt und auch als Panelisten klassifiziert (S^{plus}). Dazu wurde die Teilnahmewahrscheinlich in Abhängigkeit der vereinfachten Supermarktpräferenz konstruiert. Die Teilnahmewahrscheinlichkeit soll also auch von der interessierenden Variable abhängen.

Abweichend von der tatsächlichen durchschnittlichen Teilnahmewahrscheinlichkeit von 17 % wurden allen Befragten eine Teilnahmewahrscheinlichkeit von 30 % zugeschrieben, bis auf die Gruppe derjenigen, die die Migros präferieren. Bie dieser Gruppe beträgt die Teilnahmewahrscheinlich-

keit lediglich 10%. Tabelle 9.13 fasst dies nochmals zusammen.

	Anteil in GG^{816}	Anteil in S^{alt}	Konstruierte Teilnahmewahrscheinlichkeit	Anteil in S^{plus}
Migros	62.9%	58.9%	0.1	48.3%
Coop	26.3%	28.4%	0.3	37.4%
Denner	3.4%	4.9%	0.3	6.1%
Sonstige	7.5%	7.8%	0.3	8.1%

Tabelle 9.13: Konstruktion von S^{plus}

Die neue Stichprobe umfasst 242 Befragte statt der eigentlichen 138. Sie enthält entsprechend der Konstruktion einen hohen Anteil von Migros-Kunden.

Neue Modellierung

Um die neue Teilnahmewahrscheinlichkeit zu bestimmen, musste wieder neu modelliert werden. Das Vorgehen entspricht dem aus Kapitel 8, ab S. 137: Ausgehend von einem Modell mit allen plausiblen Variablen wurde eine Modellreduktion vorgenommen. Dabei wurden allerdings keine Einschränkungen (wie insbesondere einer Höchstzahl an Variablen) getroffen. Nichtsdestoweniger hat sich die Anzahl von Variablen auf drei reduziert. Tabelle 9.14 listet alle Variablen mit ihren Faktorenniveaus auf.

	Estimate	Std. Error	z value	Pr(> |z|)	
(Intercept)	0.1966	0.3572	0.55	0.5821	
`f851004`	0.4872	0.3850	1.27	0.2057	
`f851005`	0.7727	0.3639	2.12	0.0337	*
`f851006`	0.6615	0.3537	1.87	0.0614	.
`noNA`	-0.5159	0.0546	-9.44	0.0000	***
`f03001`	-1.1501	0.1747	-6.58	0.0000	***
Signif. Niveau: 0 *** 0.001 ** 0.01 * 0.05 .					

Tabelle 9.14: Modellierung der Teilnahmewahrscheinlichkeit mit S^{plus}

Neben der Häufigkeit das Internet zu benutzen und der Anzahl fehlender Werte (für alle Variablen) hat insbesondere die Präferenz für die Migros (`f03001`) eine hohe Erklärungskraft, was der Konstruktion der Stichprobe

entspricht.

Das Modell hat keine sehr hohe Erklärungskraft, ist aber auch nicht schlecht, wie die in Tabelle 9.15 aufgelisteten Informationskriterien zeigen:

	McF.R^2	CSR^2	Nag.R^2	AIC	BIC/SBC	κ
0.15	0.17	0.24	851	850	34	

Tabelle 9.15: Gütekriterien des Modells mit S^{plus}

Es ergeben sich Teilnahmewahrscheinlichkeiten, die gemäss Tabelle 9.16 und dem Histogramm in Abbildung 9.3 verteilt sind.

Minimum	0.02
1. Quartil	0.14
Median	0.30
Arith. Mittel	0.30
3. Quartil	0.42
Maximum	0.72

Tabelle 9.16: Verteilung der Teilnahmewahrscheinlichkeiten des Modells mit S^{plus}

Verteilung der Teilnahmewahrscheinlichkeiten

Die berechneten Teilnahmewahrscheinlichkeiten sind viel gleichmässiger Verteilt als bei den Modellen der vorangegangenen Abschnitte. Insbesondere gibt es nicht mehr eine so grosse Gruppe von Befragten mit sehr niedrigen Teilnahmewahrscheinlichkeiten. Die Informationen, die insbesondere in `f03001` (Präferenz Migros) enthalten sind, sind sehr hilfreich bei der Modellierung der Teilnahmewahrscheinlichkeiten.

Wenn man Gewichte aus den Teilnahmewahrscheinlichkeiten als einfaches Reziprok ableitet, reichen diese von 1.39 bis 35.2. Das Verhältnis von grösstem zu kleinstem Gewicht beträgt damit absolut 25.3 und bei den Befragten, die in S^{plus} sind, also den virtuellen Panelisten 13.4, denn das Maximum beträgt hier 18.6. Das sind Verhältnisse, die vertretbar sind; die Gewichte müssen nicht trunkiert werden.

Bildet man die Gewichte auf Grund klassifizierter Teilnahmewahrscheinlichkeiten, mit in diesem Fall fünf Klassen, die bei den Panelisten ungefähr gleich gross sind, ergeben sich die fünf Gewichte und die entsprechende Häufigkeit gemäss Tabelle 9.17:

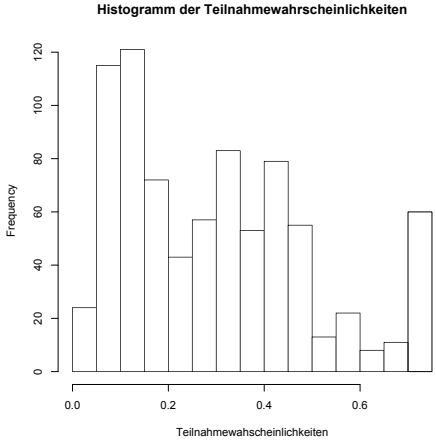


Abbildung 9.3: Histogramm der Teilnahmewahrscheinlichkeiten im Modell mit S^{plus}

Gruppe	Grösse in GG^{816}	Grösse in S^{plus}	gerundetes Gewicht
1	79	54	0.43
2	166	79	0.62
3	56	24	0.69
4	151	38	1.18
5	364	47	2.30

Tabelle 9.17: Klassifizierte Gewichte in S^{plus}

Das Verhältnis von grösstem zu kleinstem Gewicht beträgt 5.3 und ist damit relativ klein, also unproblematisch. Tabelle 9.18 vergleicht die Wirkung der beiden Gewichtungsarten.

Es kann also eine deutliche (und signifikante) Verbesserung der Schätzer für die Anteile entwickelt werden, wenn eine Gewichtung vorgenommen wird. Es scheint so zu sein, dass die direkte Gewichtungsmethode besser geeignet ist, als die Klassifizierte.

	Anteil ungewichtet		Anteil gewichtet	
	in GG^{816}	in S^{plus}	direkt	kategorisiert
Migros	62.9	48.3	58.8	55.5
Coop	26.3	37.4	27.7	30.8
Denner	3.4	6.1	2.7	2.7
Sonstige	7.5	8.1	10.8	11.0

Tabelle 9.18: Vergleich der Gewichtungsarten Modell S^{plus}

9.3.2 Simulation

Um zu zeigen, dass dieses Resultat verlässlich ist, wurden eine Mini-Simulation durchgeführt, indem die zusätzlich zu den Panelisten aufgenommenen Befragten in S^{plus} wiederholt zufällig gezogen wurden. Die Anzahl der Wiederholungen beträgt 1'000.

Tabelle 9.19 listet jeweils Minimum, arithmetisches Mittel und das Maximum der Schätzer für die Supermarktpräferenz auf.

	Anteil ungewichtet		direkt gewichtet			klassifiziert gewichtet		
	in GG^{816}	in S^{plus}	Min	Mean	Max	Min	Mean	Max
Migros	62.9	48.3	56.7	62.4	68.0	52.0	59.4	67.0
Coop	26.3	37.4	21.0	27.1	33.1	22.0	29.4	37.4
Denner	3.4	6.1	0.8	1.9	4.2	0.8	2.1	4.7
Sonstige	7.5	8.1	4.5	8.6	15.7	4.9	9.1	15.2

Tabelle 9.19: Simulierte Supermarktpräferenz

Es zeigt sich klar, dass die gewichteten Anteilsschätzer besser sind als die ungewichteten Schätzer. Ausserdem scheint es besser zu sein, wenn man das reziprok der Teilnahmewahrscheinlichkeit direkt als Gewicht benutzt und nicht etwa das Gewicht, dass sich aus dem Grössenverhältnis der Klassen ableitet. Aber auch letzteres schneidet deutlich besser ab, als der nicht gewichtete Schätzer.

9.4 Empirischer R-Indikator

Für den Datensatz GG[816] kann der R-Indikator bestimmt werden (RISQ -R-Indikator). Dieser beträgt 0.779. Allerdings scheint der R-Indikator nicht das zu messen, was er messen soll! Um das Verhalten des R-Indikators bei diesem Datensatz besser zu verstehen, wurde der Response wieder versuchsweise modifiziert, indem einfach verscheidene Gruppen zu den nicht Teilnehmern gezählt wurden. Tabelle 9.20 gibt die R-Indikatoren für einen Response ohne die jeweils aufgeführte Gruppe an. Ausserdem listet sie die Grösse n_S der jeweils verbleibenden Stichprobe auf.

Bis auf das Beispiel des Ausfalls der untersten Altersklasse verbesserte sich der R-Indikator jeweils um einige Prozentpunkte. Der Standardfehler für den R-Indikator[6] ohne Ausschlüsse beträgt 0.034. Das 95 % Konfidenzintervall beträgt daher [0.712,0.845]. Ausser dem Ausschluss der Altersklasse 2 liegen

		n_S	R_{RISQ}
Alle		125	0.779
Ohne Einkommensklasse	1 & 2	109	0.799
	3	97	0.815
	4	106	0.808
	5	103	0.845
	6	113	0.787
Ohne Altersklasse	1	106	0.770
	2	56	0.897
	3	88	0.800

Tabelle 9.20: R-Indikatoren

6 Obwohl die Berechnung des Standardfehlers schon im R-Package implementiert ist, ist die die Beschreibung der genauen Berechnung erst für das Ende des RISQ Projekts angekündigt, was zum Zeitpunkt der Berechnung noch eine halbes Jahr vor dem Termin gewesen ist. Auch das R-Package befindet sich noch in einem Produktionsstatus und ist zwar auf der Seite des RISQ-Projekts veröffentlicht (`http://www.risq-project.eu/tools/RISQ_R-indicators_v1.0.r`, aber noch nicht auf dem alle R-Packages verwaltenden CRAN-Repository (`http://cran.r-project.org/web/packages/`) abrufbar und damit noch kein «offizielles» Package. Stand: 01.11.2011.

alle R-Indikatoren innerhalb dieses Konfidenzintervalls. Die Unterschiede sind also auch nicht signifikant.

Trotzdem ist es überraschend, dass sich alle R-Indikatoren in die selbe Richtung bewegt haben, nämlich grösser geworden sind. Das kann zufällig sein, aber auch inhaltlich relevante Ursachen haben. Zunächst ist es möglich, dass das Konzept des R-Indikators der RISQ Gruppe noch nicht ganz ausgereift ist. Die wahrscheinlichere Erklärung ist allerdings das weiterhin vorhandene Problem, dass die zu Grunde liegende Schätzung der Teilnahmewahrscheinlichkeit von so ungenügender Qualität ist, dass der abgeleitete R-Indikator nicht interpretierbar ist. Abgesehen davon wurde das Minimum des R-Indkators verändert, da die Teilnahmerate gesunken ist. Aussderm hat sich das Modell geändert.

10 Simulationen

► Bisher hat sich das PSA nicht als sehr hilfreiche Methode erwiesen, eine Biasreduktion vorzunehmen. Das mag verschiedene Ursachen haben, auf die noch einzugehen sein wird. Eine wichtige Ursache aber war, dass zu wenige Informationen bezüglich des Teilnahmeverhaltens vorgelegen haben, es konnte jedenfalls nicht befriedigend modelliert werden. Um die Methode des PSA selbst zunächst besser zu verstehen, wurden eine Reihe von Simulationen durchgeführt, die in diesem Kapitel besprochen werden. Dazu müssen zunächst, abgeleitet aus den Variablen der experimentellen Befragung, künstliche Variablen erzeugt werden, die mehr Informationen über das Teilnahmeverhalten enthalten, als die empirisch erhobenen Variablen. Es werden verschiedene Versionen von Simulationen vorgestellt werden. Wieder werden auch die R-Indikatoren bestimmt werden. Im letzten Abschnitt wird dann das PSA mit einer anderen Variante, Gewichte zu bstimmen, verglichen. ◄

Die Simulationen haben allgemein einen Ablaufplan gemäss Abbildung 10.1

Variablen konstruieren

Zunächst wurde eine abhängige Variable `y` und eine Responsewahrscheinlichkeit `r` konstruiert. Datengrundlage bildet weiterhin GG[816]. Um die Variablen zu konstruieren, wurden aus der experimentellen Befragung drei beliebige Variablen ausgewählt: x_1=`f03501`, x_2=`f03502` und x_3=`f03503` («Marktforschungsunternehmen behandeln die Daten vertraulich.», «Umfragen bringen Abwechslung und sind interessant.» und «Bei Umfragen wird häufig etwas gefragt, was niemand etwas angeht.»). Alle drei Variablen haben ursprünglich fünf Ausprägungen, wurden aber trichotomisiert, um sehr kleine Häufigkeiten bei einzelnen Ausprägungen zu vermeiden. Alle drei Variablen sind jetzt kodiert als (1,2,3). Tabelle 10.1 listet die Häufigkeitsverteilung für die drei Variablen auf.

Abhängige Variable `y`

Die abhängige Variable `y` wurde mit der allgemeinen Funktion

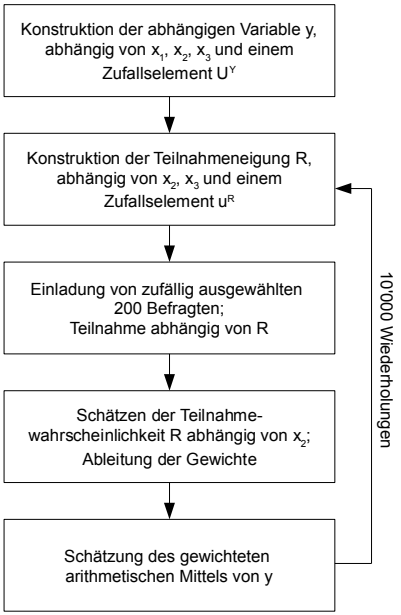


Abbildung 10.1: Ablaufplan der Simulation

$$y = a_1 x_1 + (a_2 + x_2)^2 + a_3 x_3 + u^y \tag{10.1}$$

konstruiert. y wird neben den drei transformierten Variablen aus einem Zufallsterm u^y gebildet, um die Variabilität von y zu erhöhen. Es wurden

Ausprägung	x_1	x_2	x_3
1	343	264	257
2	238	279	300
3	235	273	259

Tabelle 10.1: Häufigkeitsverteilung der Variablen x_1, x_2 und x_3

viele verschiedene Möglichkeiten ausprobiert, um die beiden Koeffizienten a_i zu bestimmen.

Ausserdem wurde eine Teilnahmewahrscheinlichkeit `r` bestimmt, die mit `y` korreliert. Allgemein hat die Vorschrift die Form

teilnahmewahrscheinlichkeit `r`

$$\text{logit}(r) = \beta_2 x_2 + \beta_3 x_3 + u^r \tag{10.2}$$

Die Teilnahmewahrscheinlichkeit `r` wird geschätzt als $\hat{r} = f(x_2)$. x_2 wurde im Vergleich zu den anderen beiden Variablen etwas verstärkt, um einen besseren Unterschied zwischen der Stichprobe und der Grundgesamtheit zu sehen. In der ersten Variante der Simulationen wird `y` konstruiert als

$$y = 5x_1 + (x_2 + 3)^2 + 5 * x_3 \tag{10.3}$$

x_1 und x_2 haben also die möglichen Ausprägungen (5,10,15), x_3 dagegen (16,25,36) Es sollen damit zwei Dinge erreicht werden: `y` und `r` sollen korrelieren, `y` soll aber mehr Informationen enthalten. Die Zufallsterme u^y und u^r sind unabhängig voneinander aus einer Normalverteilung abgeleitet, so dass der Erwartungswert 0 und die Standardabweichung 1 beträgt.

Auch bei der Konstruktion von r wurden viele verschiedene Möglichkeiten ausprobiert, bis die β endgültig festgelegt werden konnten. Wir verwenden in der ersten Variante der Simulation:

$$\text{logit}(r) = -8 + \beta_2 (x_2)^2 + 2x_3 + u^r \tag{10.4}$$

Die Subtraktion von 8 war notwendig, um die Teilnahmewahrscheinlichkeiten so zu konstruieren, dass tatsächliche Ausfallprozesse abbgebildet werden können. Da der Achsenabschnitt in anderen Modellvarianten variiert werden wird, wird diese «Variante -8» genannt. Abbildung 10.2 zeigt die Verteilung der Teilnahmewahrscheinlichkeiten für verschiedene Summanden[1].

Variante -8

Die Stichprobe wird innerhalb der Simulation so gezogen, dass jeweils eine

Auswah der Stichprobe

1 Bei den folgenden Abbildungen wurde der Zufallsterm nicht fixiert, sondern jeweils zufällig ausgewählt. Der Rekonstruierbarkeit wegen wurde allerdings der Zufallsmechanismus fixiert. Technisch ausgedrückt wurde in R `set.seed(0)` gesetzt. Bei Abbildung 10.2 (und nur bei dieser) wurde nur jede fünfte Beobachtung (163) geplottet, um eine unhandliche Grösse der Grafik zu vermeiden.

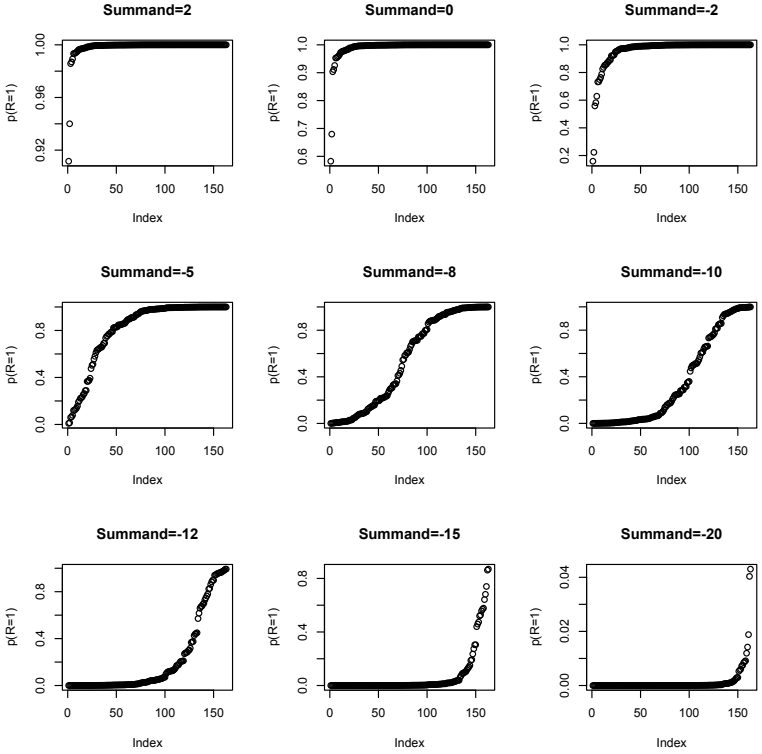


Abbildung 10.2: Mögliche Summanden in der Konstruktionsvorschrift für `r`

fixe Anzahl Personen ausgewählt werden, wobei jede eine Wahrscheinlichkeit von $r_i / \sum r$ hat, in die Stichprobe zu gelangen. Dass die Anzahl Personen nicht in Abhängigkeit von `r` schwankt, soll den Charakter von Web-Panel-Befragungen wiederspiegeln. Dort ist ist es auch üblich, dass die Anzahl der zu befragenden Personen ex ante festgelegt wird, um dann so lange weitere Personen einzuladen, bis die Vorgabe erfüllt ist. (Bei anderen Befragungsmodi ist dieses Vorgehen eher unüblich.)

Die Korrelationen zwischen den `X` Variablen und jeweils `y` und `r` ist in

Abbildung 10.3 dargestellt. Die Bildungsvorschriften spiegeln sich in den sechs Plots sehr gut wieder.

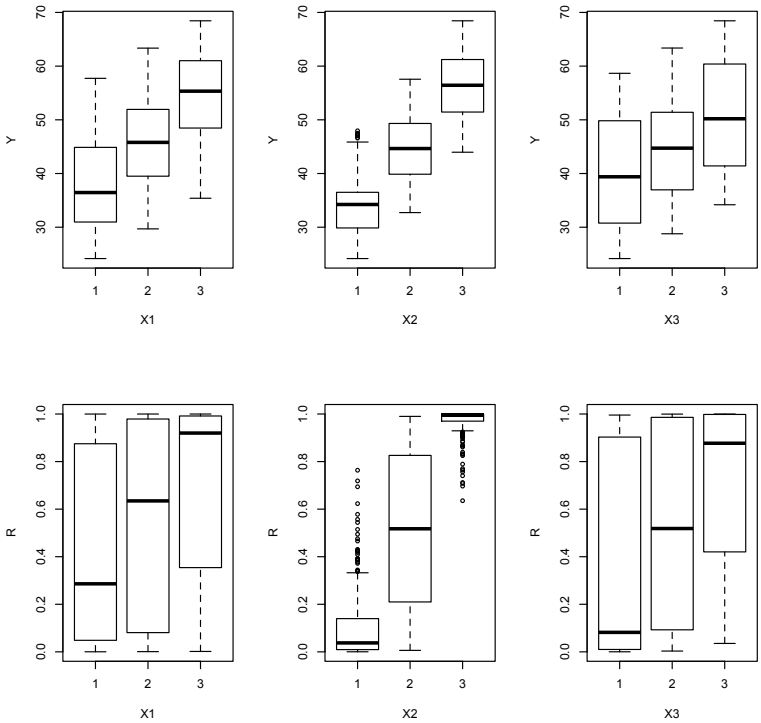


Abbildung 10.3: Boxplot von `y` und `r` gegen die sie konstruierenden Variablen x_1 bis x_1

Es ist notwendig, dass `y` und `r` korrelieren, damit es zu einem Bias beim geschätzten Mittelwert von $\hat{y}$ in der Stichprobe kommt. In Abbildung 10.4 wird `y` gegen `r` geplottet. Es besteht eine positive Korrelation von $R^2 = 0.73$. `r` ist so verteilt, dass es zu einer Häufung der Fälle an den Rändern kommt. Es gibt also einzelne Beobachtungen mit einer sehr niedrigen und einige mit einer sehr hohen Teilnahmewahrscheinlichkeit. Tendenziell ist `y` um so

grösser, je grösser auch die Teilnahmewahrscheinlichkeit ist. Es ist daher zu erwarten, dass $\bar{y}$ in einer ungewichteten Stichprobe überschätzt wird.

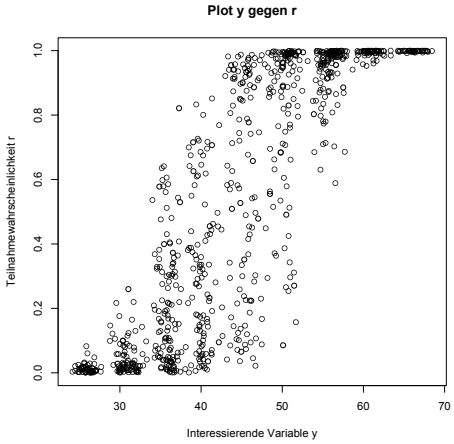


Abbildung 10.4: Plot der interessierenden Variable `y` gegen die Teilnahmewahrscheinlichkeit `r`

Schätzung von `r`

In der Simulation wird $\hat{r}$ nur mit Hilfe von x_2 geschätzt: $\text{logit}(\hat{r}) = \hat{\beta}_0 + \hat{\beta}_2 x_2$. Da neben dem Fehlerterm eine die tatsächliche Teilnahmewahrscheinlichkeit `r` bestimmende Variable fehlt und der Zusammenhang zwischen `r` und x_2 eigentlich quadratisch ist, ist die Vorhersage nicht sehr zuverlässig und ungenau. In der Realität entspricht dies einer Situation mit relativ wenigen verfügbaren Informationen über das Teilnahmeverhalten. Abbildung 10.5 zeigt einen beispielhaften Zusammenhang zwischen der konstruierten, also «echten» Teilnahmewahrscheinlichkeit und der geschätzten. Die Korrelation zwischen beiden ist mit $R^2 = 0.73$ relativ stark. Allerdings hat die geschätzte Teilnahmewahrscheinlichkeit $\hat{r}$ nur drei Ausprägungen, da auch die x_2 nur drei Ausprägungen hat.

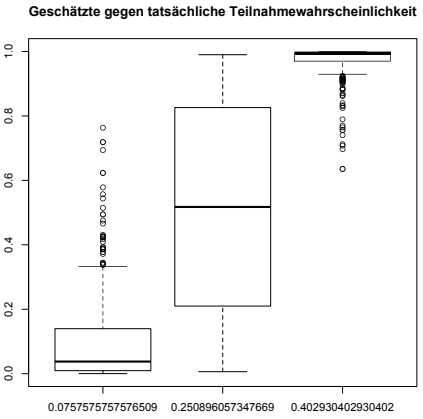


Abbildung 10.5: Plot der interessierenden Variable `y` gegen die Teilnahmewahrscheinlichkeit `r`

10.1 Simulation des PSA

10.1.1 Variante -8

Die Erklärungskraft der Modelle in den Simulationen ist weder besonders stark, noch niedrig. Der Median des Nagelkerke R^2 beträgt bei einem Minimum von 0.10 und einem Maximum von 0.28 genau wie der Mittelwert 0.18.

Tabelle 10.2 zeigt die Verteilung des geschätzten Mittelwerte von Y in den Simulationen. Der tatsächliche Mittelwert von `y` beträgt **45.11** (Standardabweichung 11.2, Minimum 24.2, Maximum 68.4).

Offensichtlich führen die gewichteten Schätzungen zu viel besseren Resultaten, als die ungewichteten! Selbst die schlechteste Schätzung mit gewichteten Schätzern führt zu einer besseren Schätzung als die beste Schätzung ohne Gewichte. Nichtsdestoweniger wird der Mittelwert auch gewichtet weiterhin überschätzt, wenngleich auch nicht so stark. Eine vollständige Eliminierung des Bias ist mindestens bei den hier verwendeten Parametern nicht möglich.

	ungewichtet	gewichtet
Minimum	50.09	45.45
Median	51.87	47.02
Maximum	53.65	48.74
Varianz	0.51	0.44

Tabelle 10.2: Mittelwerte von `y` in der Simulation

10.1.2 Variante -12

Es ist interessant herauszufinden, ob die Methode auch noch funktioniert, wenn das Teilnahmeverhalten in dem Sinne extremer wird, dass die Gruppe der Befragten mit geringer Teilnahmewahrscheinlichkeit zunimmt. Im hier gegebenen Rahmen lässt sich das dadurch am einfachsten realisieren, dass der Summand in Gleichung 10.4 zur Bestimmung der Teilnahmewahrscheinlichkeit verkleinert wird, siehe auch Abbildung 10.2 auf S. 182. Statt -8 soll er hier -12 betragen.

Abbildung 10.4 auf S. 184, in der die Teilnahmewahrscheinlichkeit `y` gegen `r` geplottet wurden, verändert sich wie in Abbildung 10.6 dargestellt. Die interesserende Variable `y` ist jetzt in der Gruppe derer mit mit einer hohen Teilnahmebereitschaft (die deutlich kleiner geworden ist) jetzt deutlich grösser als bei denjenigen mit kleiner Teilnahmewahrscheinlichkeit. Der zu erwartende Bias sollte also deutlich stärker sein, als in der Variante mit einem Summanden von -8.

Die Erklärungskraft der Modelle in den Simulationen hat deutlich zugenommen. Der Median des Nagelkerke R^2 beträgt bei einem Minimum von 0.40 (vorher 0.10) und einem Maximum von 0.64 (0.28) jetzt im Median 0.53 (0.18). Diese deutliche Zunahme ist zunächst überraschend, kann aber vermutlich damit erklärt werden, dass der Einfluss von x_2 bei der Konstruktion von `r` zugenommen hat, wie der Vergleich der Boxplots in Abbildung 10.3, auf S. 183 mit Abbildung 10.7 zeigt. (Da sich an der Konstruktion von `y` nichts geändert hat, sind nur die Boxplots `r` gegen x_i aufgeführt). Ist bei der Variante -8 der Median von `r` bei $x_2 = 2$ noch sehr nahe bei Null, ist er jetzt deutlich höher.

Die folgende Tabelle 10.3 vergleicht zeigt die Verteilung des geschätzten

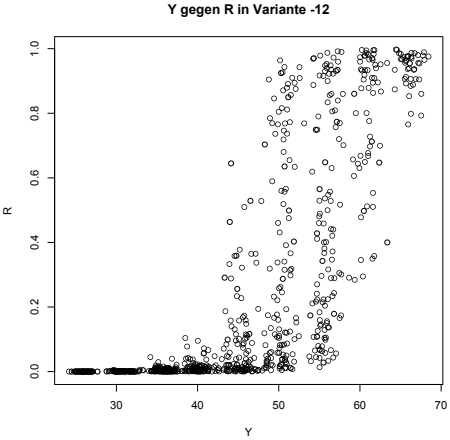


Abbildung 10.6: Plot der interessierenden Variable `y` gegen die Teilnahmewahrscheinlichkeit `r`

Mittelwerte von `y` in den Simulationen. Der tatsächliche Mittelwert von `y` beträgt wie auch schon in der Variante −8 weiterhin **45.11** (Standardabweichung 11.2, Minimum 24.2, Maximum 68.4).

	ungewichtet	gewichtet
Minimum	55.19	43.43
Median	56.52	48.26
Maximum	57.64	54.92
Varianz	0.31	2.16

Tabelle 10.3: Mittelwerte von `y` der Simulation in Variante −12

Wie zu erwarten, ist der Bias im Vergleich zu Variante −8 (siehe Tabelle 10.2 auf S. 186) deutlich stärker geworden. Der Median beträgt jetzt ungewichtet 56.52 im Vergleich zu 51.87. Es ist trotzdem noch eine deutliche Biasreduktion mittels Gewichtung möglich. Der Median des Schätzters liegt jetzt zwar nicht mehr ganz so nahe am tatsächlichen Wert, wie in Variante −8, allerdings ist er nicht viel schlechter.

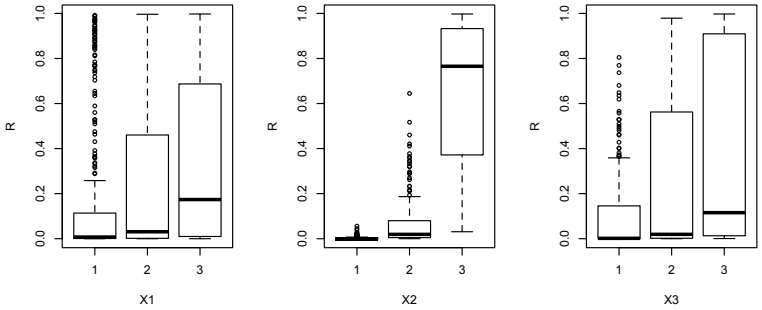


Abbildung 10.7: Boxplot von r gegen die konstruierenden Variablen x_1 bis x_1 in Variante -12

Allerdings hat die Varianz des Schätzers deutlich zugenommen. Bezüglich des Erwartungswertes ist es zwar dringend geraten, den gewichteten Schätzer zu verwenden, allerdings ist dieser auch variabler. Allerdings ist es immer noch so, dass bei 10'000 Simulationsdurchläufen die schlechteste gewichtete Schätzung besser war, als die beste ungewichtete, wenn auch nur knapp.

10.1.3 Variante -20

In einer weiteren Variante wurde das Teilnahmeverhalten als noch extremer simuliert. In Gleichung 10.4 zur Bestimmung der Teilnahmewahrscheinlichkeit wurde der Summand jetzt auf -20 verkleinert, siehe auch Abbildung 10.2 auf S. 182.

Es gibt jetzt nur noch einige wenige Beobachtungen mit einer (relativ) hohen Teilnahmewahrscheinlichkeit. Diese Beobachtungen werden in den Simulationen mit grosser Wahrscheinlichkeit in die Stichprobe gelangen und sollten einen entsprechend hohen Einfluss auf den Bias haben, siehe Abbildung 10.8.

Die Erklärungskraft der Modelle hat nochmals zugenommen, der Median beträgt jetzt 0.63. Wiederum liegt die Vermutung nahe, dass dies durch einen gestiegenen Einfluss von x_2 zurückzuführen ist.

Der Bias ist auch tatsächlich nochmals gestiegen, wobei immer noch eine

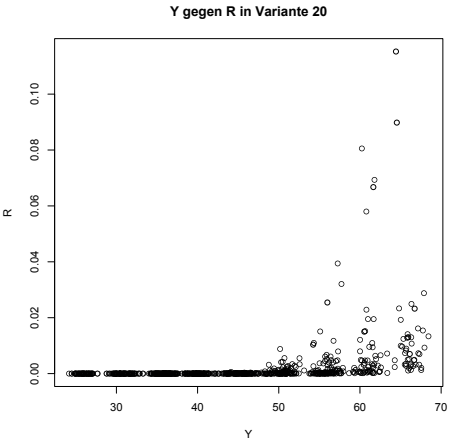


Abbildung 10.8: Plot der interessierenden Variable y gegen die Teilnahmewahrscheinlichkeit r in Variante -20

deutliche Verbesserung der Schätzung durch Gewichtung möglich ist, siehe auch Tabelle 10.4.

	ungewichtet	gewichtet
Minimum	57.00	42.81
Median	57.80	52.18
Maximum	58.50	56.40
Varianz	0.31	2.82

Tabelle 10.4: Mittelwerte von y der Simulation in Variante -20

Weiterhin ist es so, dass der gewichtete Schätzer bei 10'000 Durchläufen immer besser ist als der ungewichtete. Die Varianz hat weiterhin zugenommen. Ausserdem ist der gewichtete Schätzer jetzt deutlich schlechter als in den vorangegangenen Simulationsvarianten und hat sich also mehr vom tatsächlichen Mittelwert von y entfernt.

Das Verhältnis vom grössten zum kleinsten Gewicht hat sich stark geändert. Die folgende Tabelle 10.5 listet die Verhältnisse für die bisher beschrie-

benen Simulationsdurchläufe auf:

	Variante-8	Variante-12	Variante-20
Minimum	3.6	6.9	28.7
Median	6.9	19.6	206 Mio.
Maximum	24.2	98 Mio.	226 Mio.

Tabelle 10.5: Maximale Gewichtsverhältnisse in den Simulationsdurchläufen

Die Verhältnisse sind in Version -20 so extrem, dass wenn man nicht mittels Simulationen weiss, dass tatsächlich die Schätzer verbessert wurden, man sicher auf eine Gewichtung verzichten würde. Dies gilt auch für einige Fälle in Variante -12.

10.1.4 Variante -20ex

Als letzte Variante soll überprüft werden, wie sich der Schätzer verhält, wenn die Teilnahmewahrscheinlichkeit nicht mehr gut geschätzt werden kann. Daher wurde die Konstruktion der Variablen nochmals modifiziert. Damit weiterhin ein Bias auftritt, wurde der Einfluss von x_3 sowohl bei der Konstruktion von y wie auch r erhöht. Es gilt jetzt:

$$\text{logit}(r) = -20 + x_2^2 + 10x_3 + u^r \tag{10.5}$$

und

$$y = 5x_1 + (x_2 + 3)^2 + 5 * x_3 \tag{10.6}$$

x_3 wurde deswegen verstärkt, weil es sowohl in der Konstruktion von y und r verwendet wird, nicht jedoch bei der Schätzung von $\hat{r}$.

Es gehen also im Vergleich zu den vergangenen Modellen Informationen zur Modellierung der Teilnahmewahrscheinlichkeit verloren. Der zusätzliche Summand musste relativ klein sein, da die Gesamtsumme der Gleichung relativ zu den andren Varianten hoch ist. Wäre der Summand nicht in ungefähr dieser Grösse, hätten wieder die meisten Befragten eine entweder sehr kleine oder alle eine sehr grosse Wahrscheinlichkeit, gezogen zu werden (was bezogen auf den Umstand, dass 200 Befragte sicher gezogen werden,

keinen Unterschied macht). Vergleiche auch das ähnlich gelagerte Problem in Variante -8, Abbildung 10.2, S. 182.

Die Modelle zur Prognose der Teilnahmewahscheinlichkeit sind erwartungsgemäss nicht gut. Das Nagelkerke R^2 liegt bei einem Median von 0.002 meist sehr nahe bei Null.

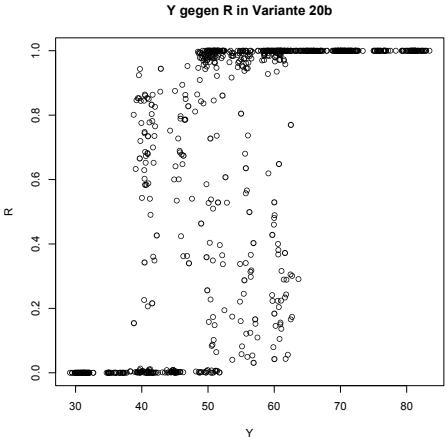


Abbildung 10.9: Plot der interessierenden Variable `y` gegen die Teilnahmewahrscheinlichkeit `r` in Variante -20ex

Wie in den vorangegangenen Varianten gibt es wieder zwei extreme Gruppen, mit sehr kleinen und sehr grossen Teilnahmewahrscheinlichkeiten. Beide unterscheiden sich wieder bezüglich der interessierenden Variable `y` so, dass wieder ein Bias zu erwarten ist, falls nicht gewichtet wird. Tabelle 10.6 zeigt, dass dies auch tatsächlich der Fall ist. Der tatsächliche Mittelwert von `y` beträgt so konstruiert **55.13**

Es ist keine Überraschung, dass die Verbesserung der Schätzer bei weitem nicht mehr so gut ausfällt, wie in den vorangegangenen Varianten. Wenn man allerdings bedenkt, wie schlecht die Teilnahmewahrscheinlichkeit nur geschätzt werden konnte (sehr niedrige R^2), ist es allerdings doch erstaunlich, dass eine Verbesserung der Schätzer erreicht werden konnte. Sogar die Varianz des gewichteten Schätzers ist kleiner als der des ungewichteten.

	ungewichtet	gewichtet
Minimum	57.91	57.63
Median	60.48	59.31
Maximum	62.95	60.85
Varianz	0.63	0.42

Tabelle 10.6: Mittelwerte von `y` der Simulation in Variante −20ex

10.2 R-Indikatoren und Gewichtung

In einem weiteren Durchlauf der Simulationen mit 10'000 Wiederholungen wurde getestet, ob die Güte der Gewichtung mittels R-Indikator vorhergesagt werden kann. In der Praxis besteht das Problem, dass nicht klar ist, ob eine Gewichtung mittels Teilnahmewahrscheinlichkeit erfolgversprechend ist. Eine mögliche Hilfe ist die Anwendung der in Abschnitt 3.5.2 (ab S. 47) beschriebenen R-Indikatoren. Ausserdem bietet der Vergleich beider R-Indikatoren nochmals die Möglichkeit, GREG und PSA zu vergleichen, da der R-Indikatoren von Särndal und Lundström (2008) (im Folgenden kurz `R`) GREG-basiert und der RISQ-Gruppe (`Q`)[2] PSA basiert ist.

Grundlage waren die Simulationsparameter gemäss Variante -8 der vorhergehenden Simulationen, siehe Abschnitt 10.1.1. Es gelten daher Gleichung 10.4 zur Bestimmung der Teilnahmewahrscheinlichkeit und Gleichung 10.3 zur Definition der interessierenden Variable `y`. Die Bestimmung beider R-Indikatoren stützt sich auf die Variablen x_1 und x_2 ab.

10.2.1 Bestimmung der R-Indikatoren

`R` RISQ selbst stellt den Quell-Code zur Berechnung des Indikators zur Verfügung[3]. Dieser Code wurde zwar übernommen und in die Simulationen implementiert, war aber im Resultat nicht sehr überzeugend. Bei 7'101 der

2 Die Notation bzw. entspricht der Wahl der RISQ-Gruppe, z. B. Shlomo et al. (2009).

3 Zum Zeitpunkt der Durchführung der Simulationen existiert das angekündigte Packet für R (noch?) nicht. Allerdings stehen Teile des Quell-Codes für dieses Packet bereits zur Verfügung: `http://www.risq-project.eu/tools/RISQ_R-indicators_v1.0.r`. Ausserdem ist die Implementation beschrieben in de Heij et al. (2010).

10'000 Simulationsdurchläufe gilt R=1. Das ist nicht sehr befriedigend, da das Resultat nur dadurch bedingt ist, dass die Teilnahmewahrscheinlichkeit nicht richtig geschätzt werden konnte. Es wurde daher die Bestimmung von R mit eigenem Code nachgebaut.

Mit Hilfe des R-Packetes survey (Lumley, 2010)[4] können GREGs bestimmt werden. Das Packet erlaubt die Extraktion der sogenannten *g-weights*. Q wird hier bestimmt als Varianz der *g-weights*. Q verhält sich damit invers zu R. Die notwändigen Popuationstotale für x_1 und x_2 sind bei den Simulationen bekannt, vergleiche Tabelle 10.1. Als Stichprobenplan wurde eine einfache Zufallsauswahl ohne Zurücklegen angenommen. Als Startgewichte erhalten alle Elmente den selben Wert, berechnet als Quotient aus Stichprobengrösse und Grösse der Population, also $200/816 \approx 0.245$. Q

Das Nagelkerke R^2 ist bei einem Durchschnitt von 0.0037 und einem Maximum von 0.035 sehr niedrig. Es ist daher keine Überraschung, dass R hoch ist: wenn die Kovariate (in diesem Fall x_1 und $x_2$die Teilnahmebereitschaft nicht gut schätzen kann, heisst das, dass sich die Kovariate zwischen Teilnehmenden und Verweigerern im Sinne der Simulation nicht unterscheiden. Daher zeigt R, dem dieselben Kovariate zu Grunde liegen, auch keinen Bias an. Die Interpretation von Q ist weniger einfach. Dieser Indikator hilft vor allem, die Güte von Stichproben (unter der Voraussetzung konstanter Kovariate) zu vergleichen. Der absolute Wert ist nicht gut interpretierbar. Eigentlich gilt dies auch für R, da dieser Indikator aber auf das Interval [0,1] beschränkt ist, wobei 1 der Wert für die höchste Repräsentativität ist, ist er intuitiver. Beide R-Indikatoren sind nicht dazu gedacht, eine absolute Angabe zur Repräsentativität zu liefern, sondern sind nur bezüglich vergleichbarer Stichproben und bezüglich des gleichen zu Grunde liegenden Modells interpretierbar. Sie sind also insbesondere geeignet, die Veränderung der Repräsentativität vergleichbarer Stichproben zu überwachen, was sich insbesondere im Kontext von Panels eignet. Modellgüte

In Abbildung 10.10 ist das Nagelkerke R^2 gegen R geplottet[5] Der Zu-

4 Die Ergebnisse wurden mit denen des Package sampling von Tillé und Matei (2009) verglichen und bestätigt.

5 Bei den Plots handelt es sich um einen Dichte-Plot. Ein solcher Plot ist einem gewöhnlichen Scatter-Plot verwandt, mit dem Unterschied, dass die Dichte mittels Farbabstufungen dargestellt wird. Je dunkler die Farbgebung, desto höher die Dichte, deste mehr Beobachtungen liegen also in einem Bereich. Da die Dichte bei den vielen Beobachtungen allgemein sehr hoch ist, ergeben sich im Scatter-Plot nur schwarze Flächen.

sammenhang ist offensichtlich sehr stark, beide Grössen korrelieren mit −0.64. In Abbildung 10.11 ist das Nagelkerke R^2 gegen Q geplottet. Der Zusammenhang zwischen dem Nagelkerke R^2 und Q ist ähnlich stark wie mit R, die Korrelation beträgt 0.69. Die Korrelation zwischen R und Q ist mit -0.91 sehr hoch, wenngleich Abbildung 10.12 zeigt, dass beide Werte nicht immer dasselbe anzeigen. Insbesondere gibt es eine grosse Gruppe von Simulationsdurchläufen, bei der R=1 oder zumindest sehr nahe bei 1 ist. Q ist in diesen Fällen deutlich differenzierter.

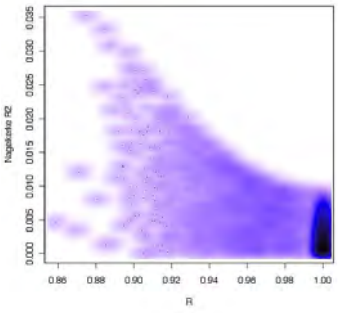


Abbildung 10.10: Plot R gegen Nagelkerke R^2

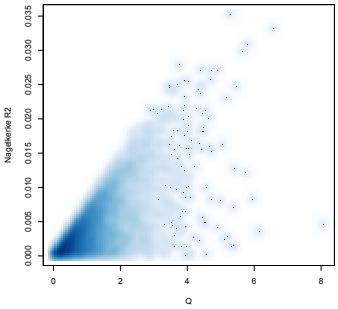


Abbildung 10.11: Plot Q gegen Nagelkerke R^2

Interessant ist, inwiefern die Indikatoren einen Bias anzeigen können.

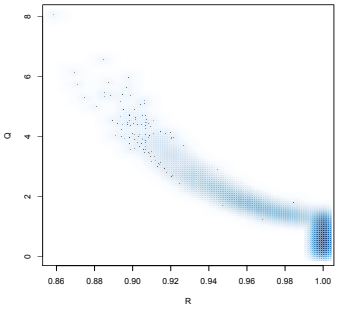


Abbildung 10.12: Plot Q gegen R

Als Indikator für den Bias soll hier einfach die Abweichung des Mittelwerts von y in der Stichprobe von dem in der Grundgesamtheit ($\bar{y}_{GG} - \bar{y}_S$ dienen. Abbildung 10.13 zeigt die Plots der Abweichungen jeweils ungewichtet gegen die beiden R-Indikatoren und nochmals die PSA-gewichteten Mittelwerte. Es zeigt sich, dass keiner der beiden Indikatoren tatsächlich eine Abweichung anzeigen kann! Beide Indikatoren sind also im Setting dieser Simulationen nicht gut als Indikatoren für einen Bias brauchbar. Sicher ist das Ergebnis anders, wenn der Zusammenhang zwischen der interessierenden variable und dem Teilnahmeverhalten noch enger konstruiert wird. Vermutlich handelt es sich bei dem hier als relativ schwach modellierten Zusammenhang empirisch um einen Extremfall, aber ein Zweifel an der Güte der Indikatoren bleibt.

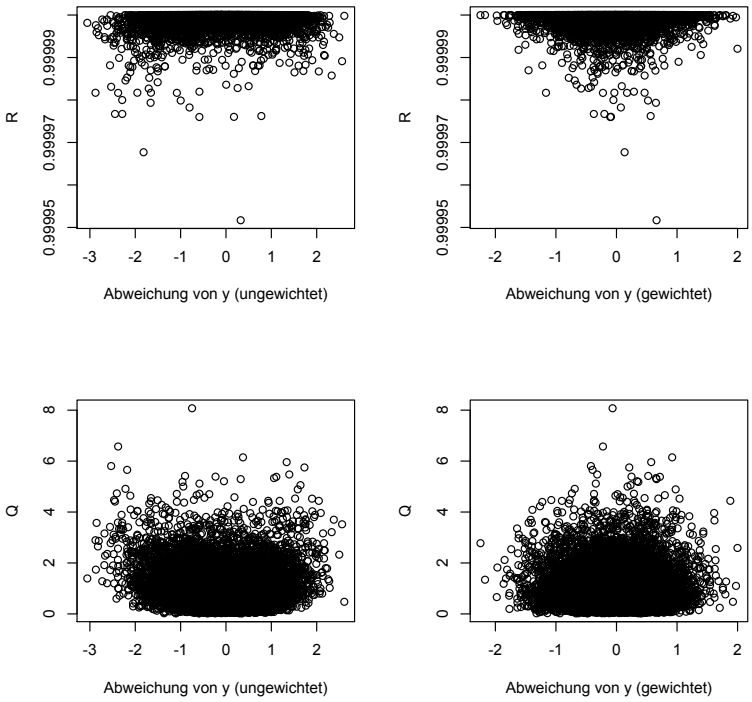


Abbildung 10.13: Plot Q und R gegen Abweichung von $\bar{y}$ gewichtet und ungewichtet

10.3 Vergleich PSA und GREG

Als Abschluss der Simulationen wurde die Schätzung mittels PSA mit dem verwandten Generalized Regression Estimator (GREG) verglichen. Zum GREG siehe auch Abschnitt 2.3 insbesondere ab Gleichung 2.9 auf S. 22.

Der Vergleich ist aus zwei Gründen interessant. Zunächst ist der GREG-Schätzer ein sehr gebräuchlicher Schätzer, der sehr häufig, insbesondere auch in der amtlichen Statistik verwendet wird. Ausserdem ist er gut softwaretechnisch umgesetzt, was die einfache Anwendung ermöglicht. In R ist er

zum Beispiel im Package `sampling` (Tillé und Matei, 2009) umgesetzt. Ausserdem ist der Vergleich des PSA mit dem GREG ein guter Test der Methode: der GREG verwendet zur Kallibrierung nur als externe Informationen die Randsummen und ist damit in der Praxis viel unkomplizierter anwendbar, da er keine Informationen über die Grundgesamtheit auf der Individualebene benötigt.

Der Vergleich soll hier auf einer relativ simple Weise erfolgen in dem Sinne, dass nicht alle möglichen Verfeinerungen des GREG getestet werden sollen. Ausserdem wird eine Vielzahl möglicher Situationen nicht verglichen. Es wird nicht untersucht, welcher Schätzer die besseren Resultate liefert, wenn die Verteilung der teilnahmewahrscheinlichkeiten auf irgendeine Art extrem wird, z. B. wenn es einzelne extreme Beobachtungen gibt, die insbesondere den GREG-Schätzer beeinflussen können und mindestens bei der klassifizierten Version des PSA weiniger einflussreich sein sollten. Ausserdem wird darauf verzichtet zu zeigen, wie sich die einzelnen Schätzer verhalten, wenn die Verteilung der Teilnahmewahrscheinlichkeiten sehr schief ist, also z. B. wenn die Teilnahmewahrscheinlichkeit so ist, dass viele Befragte nur mit sehr geringer Wahrscheinlichkeit an einer Befragung partizipieren und nur wenige eine hohe Teilnahmewahrscheinlichkeit haben.

Der GREG-Schätzer wurde nur in seiner einfachsten Form verwendet: es wurden keine *stages* mit gleichen Gewichten gebildet, die Gewichte wurden nicht getrimmt, der GREG wurde nicht robustifiziert. Es geht bei den Simulationen lediglich um den Vergleich des PSA mit dem einfachen GREG, um einen ersten Eindruck zu bekommen.

10.3.1 Referenzsimulation GREG

Konstruiert man die Teilnahmewahrscheinlichkeit wie in den vorangegangenen Simulationen mit Hilfe diskreter Variablen, die ja in der hier verwendeten Variante nur drei Ausprägungen haben, führen GREG und PSA zu genau identischen Gewichten. Das ist also kein interessanter Fall. Um die Unterschiede etwas genauer zu sehen, wurden sowohl die abhängige Variable wie auch die Teilnahmewahrscheinlichkeit etwas komplizierter konstruiert. Sie basieren jetzt auf mehr Variablen, die auch stetig sein können.

Die abhängige Variable und die Teilnahmewahrscheinlichkeit wurden jetzt aus vielen Variablen gebildet, um den besser modellieren zu können. Neben

x_1 =f03501, x_2 =f03502, x_3 =f03503 und x_4 =f03505 wurden auch x_5 =f03305 und x_6 =f03306 verwendet. Zu allen Variablen wurde jeweils einzeln ein normalverteilter Zufallsterm (Mittelwert 0, Standardabweichung 1) addiert, um stetige Variablen zu erhalten.

Die Bildungsvorschift für die Teilnahmewahrscheinlichkeit und die abhängige Variable lauten jetzt

$$\text{logit}(r) = x_2^2 + 2x_3 - (x_4 + x_5 + x_6)/2.6 + u^r \tag{10.7}$$

und

$$y = 5x_1 + (x_2 + 3)^2 + 5 * x_3 + (x_4 + x_5 + x_6)/2 + u^y \tag{10.8}$$

Die Division durch 2.6 in Gl. 10.7 ist notwendig, damit nicht alle Teilnahmewahrschinlichkeiten entweder sehr nahe bei 1 oder 0 sind. Die Verteilung sieht jetzt typischerweise aus wie in Abbildung 10.14. Es handelt sich also annähernd um eine U-förmige Verteilung, was (vermutlich) nicht sehr unrealistisch bei vielen Befragungen ist.

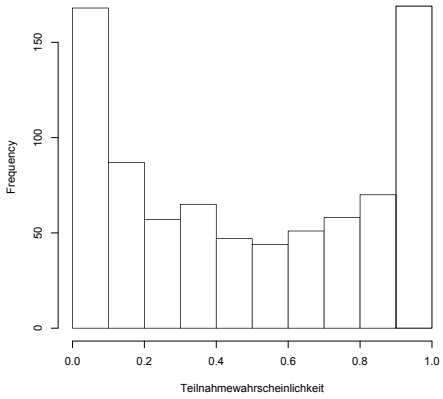


Abbildung 10.14: Verteilung der Teilnahmewahrscheinlichkeiten bei der GREG Simulation

Wieder wurden 10'000 Simulationsdurchläufe durchgeführt, wobei der Zufallsterm u^r jeweils neu bestimmt wurde. Ein weiteres Zufallselement ist jeweils neu gezogene Stichprobe. Die interessierende Variable `Y` wurde jeweils mit Hilfe eines GREG und mittels PSA geschätzt, wobei die Gewichte sowohl mittels 5er Klassifikation wie auch mit der direkten Methode berechnet wurden. Zum Schätzen werden die Variablen x_2, x_4, x_5 und x_6 verwendet. Die Teilnahmewahrscheinlichkeit `r` wurde so konstruiert, dass die zur Schätzung verwendeten Variablen die Teilnahmewahrscheinlichkeit nicht sehr gut prognostizieren können. Das Nagelkerke R^2 schwankt bei einem Durchschnitt von 0.08 zwischen 0.002 und 0.173. Zumindest das niedrige Niveau sollte wiederum realistisch sein.

Der Mittelwert von `y` beträgt in der fiktiven Grundgesamtheit 56.4. In Abbildung 10.15 werden die resultierenden Schätzer für den Mittelwert von `Y` miteinander verglichen. S bezeichnet die ungewichteten Schätzer aus der Stichprobe. Die rote Linie ist der tatsächliche Mittelwert der Grundgesamtheit.

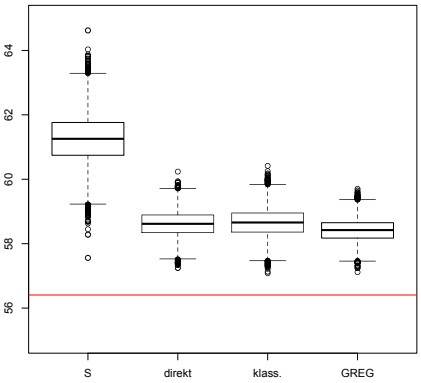


Abbildung 10.15: Referenzsimulation GREG

Alle drei Schätzer, die Gewichtungen verwenden, sind deutlich besser als die ungewichtete Schätzung. Die Unterschiede zwischen den verschiedenen Schätzern ist nicht sehr gross. Der Median des GREG-Schätzers liegt

allerdings am nächsten am tatsächlichen Wert der Grundgesamtheit. Ausserdem ist die Varianz des GREG-Schätzers am niedrigsten: $\sigma^2_{GREG} = 0.126$, $\sigma^2_{klass.} = 0.196$ und $\sigma^2_{direkt} = 0.16$. Die Varianz des ungewichteten Schätzers ist am höchsten: $\sigma^2_S = 0.589$.

Mindestens unter den Bedingungen der Simulation schneidet der GREG-Schätzer also am besten ab! Obwohl die Informationen auf Individualebene präzisere Gewichtungen ermöglichen sollten, führen sie nicht zu einem besseren Schätzer. Es ist auch bemerkenswert, dass kein Schätzer in keinem Simulationsdurchlauf den tatsächlichen Wert trifft.

10.3.2 Simulation mit höherer Korrelation

Um das Verhalten auch in anderen Situationen besser zu verstehen, wurde die Teilnahmewahrscheinlichkeit nochmals so modifiziert, dass die Kovariate eine bessere Prognose erlauben. Es wird also die Situation simuliert, dass andere Kovariate gewählt werden, die zu einer besseren Schätzung der Teilnahmewahrscheinlichkeit führen.

Gleichung 10.2 wurde so modifiziert, dass $\beta_3 = 0.2$ gilt. Der Einfluss der Variable, die zwar Teil der Bildungsvorschrift ist, aber nicht zu den Variablen gehört, die zur Prognose herangezogen werden, wird deutlich vermindert. Das resultierende Nagelkerke R^2 hat jetzt bei einem Median von 0.34 einen Bereich von 0.21 und 0.50.

Abbildung10.16 zeigt, dass sich an der Relation der Schätzer zueinander nichts geändert hat. Allerdings ist ihr Abstand zum tatsächlichen Wert kleiner geworden, was nicht überrascht.

10.3.3 Simulation GREG mit veränderter Stichprobengrösse

In zwei weiteren Simulationsdurchläufen wurde überprüft, ob die Schätzer möglicherweise auf eine Veränderte Grösse der Stichprobe reagieren und zu Unterschieden zwischen den Schätzern führen können.

Die Stichprobengrösse wurde von 200 auf 70 verkleinert und in einem weiteren Durchlauf auf 500 vergössert. Abbildung 10.17 und 10.18 zeigen, dass die Stichprobengrösse lediglich einen Einfluss auf die Varianz der Schätzer hat, nicht aber auf die Güte der Biaskorrektur. Insbesondere wenn die Stichprobe vergössert wird, unterscheiden sich die gewichteten Schätzer

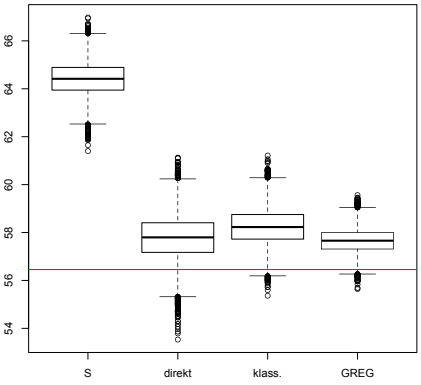


Abbildung 10.16: Simulation GREG mit höherer Korrelation

kaum noch voneinander. Ausserdem liegen sie näher am eigentlichen Mittelwert der Grundgesamtheit, was auch nicht überrascht, werden ja rund 5/8 der Grundgesamtheit befragt.

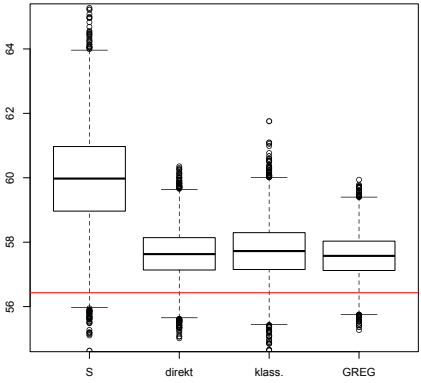


Abbildung 10.17: Simulation GREG mit verkleinerter Stichprobengrösse $n = 70$

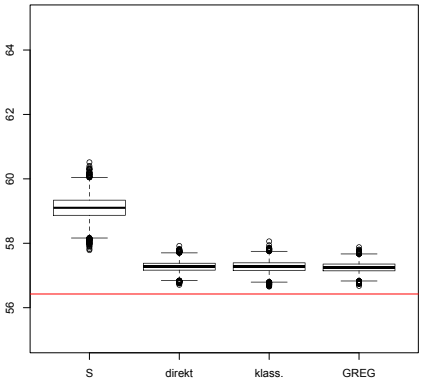


Abbildung 10.18: Simulation GREG mit vergrösserter Stichprobengrösse $n = 500$

11 Zusammenfassung und Ausblick

Es lässt sich zusammenfassen, dass eine die Ableitung von Gewichten, die mittels der Schätzung der Wahrscheinlichkeit an der Befragung teilzunehmen, nicht geeignet ist, als automatisierte Methode bei Befragungen angewendet zu werden.

PSA nur in Simulationen erfolgreich

Die Wirkung der Biaskorrektur durch die Anwendung einer solchen Gewichtung konnte nicht unmittelbar empirisch gezeigt, sondern nur mittels Simulationen nachvollzogen werden. Die Empirie ist nicht zuletzt an dem Umstand gescheitert, dass es mindestens bei den Fragen, die hier in der Hoffnung implementiert wurden, besonders gute Indikatoren für einen Bias zu sein, kaum Unterschiede zwischen Freiwilligen und der der Grundgesamtheit gegeben hat. Möglicherweise ist der Ausfallprozess zwischen der CATI Rekrutierung und der Web-Befragung nicht biasbehaftet gewesen. Man kann schlussfolgern, dass mindestens wenn ein Web-Panel sorgfältig rekrutiert wurde, es das Potential hat Daten zu liefern, die qualitativ denen aus CATI Befragungen nicht nachstehen. Mindestens für einige Variablen konnte das gezeigt werden, wobei natürlich nicht klar ist, dass das Ergebnis auch für andere Variablen gilt. Aber der geringe Unterschied ist sicherlich das interessanteste und überraschendste Ergebnis dieser Arbeit.

Aber auch die existierenden kleinen, zwar nicht signifikanten aber doch vorhandenen Unterschiede konnten nicht mittels PSA korrigiert werden. Teilweise waren die gewichtung sogar kontraproduktiv: Die Theorie sagt zwar voraus, dass sich die gewichteten Schätzer eigentlich neutral verhalten sollten, wenn die Teilnahmewahrscheinlichkeit nicht mit der interessierenden Variable korreliert, war dem empirisch aber nicht so. Die gewichteten Schätzer waren sogar noch mehr verzerrt, als die ungewichteten. In diesem Licht ist von einem PSA abzuraten, mindestens mit den hier verwendeten Variablen.

Die nämlich könnten ein **erster** Grund dafür gewesen sein, dass die Gewichtung empirisch nicht funktioniert hat. Die Simulationen konnten

zeigen, dass meistens schon geringste Informationen über den Ausfallprozess ausreichen, um eine wirksame Biaskorrektur zu ermöglichen. Das ist ein guter Beweis für das starke Potential der Methode. Vielleicht waren aber einfach die Variablen nicht in der Lage, auch nur diese Mindestanforderung an Erklärungskraft zu erfüllen. Zugegebenermassen ist dies aber unwahrscheinlich, da eine grosse Anzahl von variablen zur Verfügung gestanden hat, die wenngleich sie vielleicht auch nicht selbst das Teilnahmeverhalten gut erklären konnten, doch mindestens indirekt irgendwie mit den eigentlichen ursächlichen Variablen korrelieren sollten.

Trotzdem ist es natürlich möglich, dass die Mechanismen, die zu Response oder Nonresponse führen, im Sinne einer Handlungstheorie noch nicht verstanden wurden. Ein abschliessendes Urteil ist allerdings nicht möglich, da die Unterscheide einfach zu klein waren, um zu verstehen, ob die vorgeschlagenen Variablen die Teilnahme gut prognostizieren können.

Unabhängig davon ist aber klar, dass auch in diesem Bereich noch mehr Forschungsarbeit jenseits reiner Deskription geleistet werden muss, wenn eine Biaskorrektur erfolgreich sein soll. Nur wenn die Ausfallmechanismen besser verstanden und messbar gemacht werden, ist eine gute Biaskorrektur möglich. Diese Arbeit deutet darauf hin, dass die Theorie rationaler Wahl gute Ableitungen von Determinanten erlaubt.

Möglicherweise war **zweitens** auch die Art der Modellierung nicht angemessen. Allerdings wurden verschiedene, auch in der Literatur vorgeschlagene Möglichkeiten ausprobiert, die alle zu ähnlichen Ergebnissen geführt haben. Neben der naheliegenden Modellierung via Regressionen wurde auch Bäume ausprobiert.

Eine **dritte** Möglichkeit, warum das PSA nicht gut funktioniert hat, ist eine schlechte Datenlage. Vielleicht sind bei der Erhebung der Daten unentdeckte Fehler passiert, was ja immer möglich sein kann. Beispielsweise könnte es der Fall sein, dass viele auf die Einladung zur Web-Befragung nach der Zusage zur Teilnahme nicht reagiert haben, weil die angegebene E-Mail Adresse falsch war. Es ist nicht zwangsläufig so, dass das immer mit technischen Hilfsmitteln erkannt werden kann. Oder es sind Fehler bei der Kodierung oder Übertragung geschehen. Das sind nur Beispiele für eine Vielzahl möglicher Fehlerquellen, die ex post nicht mehr überprüfbar sind. Ratsam –aber nach jetzigem Stand der Dinge unrealistische– wäre es jedenfalls, die Untersuchung nochmals mit Hilfe neuer Daten zu replizieren.

In Abschnitt 3.2.2 ab S. 31 wurde gezeigt, dass dei Unterschiede zwischen webbasierten Befragungen und telefonischen Befragungen im Allgemeinen nicht sehr gross sind. Vielleicht wäre es besser gewesen (wenngleich auch im Rahmen des Projekts nicht umsetzbar), dass statt einer telefonischen eine schriftliche Befragung durchgeführt worden wäre. Dann wären mindestens die in der literatur berichteten Unterschiede grösser gewesen.

Ausfallprozess nur unzureichend erfasst

Ein grosses Manko dieser Arbeit, dass sich auch durch die Anbindung an ein KTI-Projekt ergeben hat, ist der Umstand, dass der gesamte Ausfallprozess zwischen der Grundgesamtheit und der eigentlichen Web-Befragung nur ausschnittsweise nachempfunden werden konnte. Es war nur möglich, den Ausfall zwischen CATI-Befragung und Web zu modellieren. Abgesehen davon, dass das Projekt so angelegt war, ist es selbstverständlich nicht trivial, sich ein erweitertes Konzept dergestalt vorzustellen, dass es möglich wäre, gesicherte Kovariate auf der Ebene der Grundgesamtheit zur Verfügung zu haben.

Hat Stoop (2005) Recht, gibt es eine Gruppe von Personen, die nur eine sehr geringe Wahrscheinlichkeit haben, generell an Befragungen zu partizipieren. Es ist sehr schwierig, Informationen über diese Gruppe zu bekommen, mindestens im Rahmen von auf Freiwilligkeit basierender Befragungen. Diese Gruppe ist nicht inferenzfähig. Denkbare wäre vielleicht, Informationen, die auf individueller Ebene aus der amtlichen Statistik zur Verfügung stehen, für eine weiterführende Korrektur heranzuziehen. Allerdings müsste es sich dann auch bestenfalls um irgendwie prozessgenerierte Daten handeln, also solche, die unabhängig von Befragungen sind. Derart sind aber typischerweise nur Registerdaten, die die meisten Variablen, die theoretisch das Teilnahmeverhalten bestimmen sollten, nicht enthalten.

Fazit: kein PSA!

Selbst wenn das Propensity Score Adjustemnt also irgendwann einmal funktionieren sollte, ist es unwahrscheinlich oder sogar unmöglich, dass es jeglichen Bias korrigieren können wird. Es wäre allerdings schön, wenn es einen Beitrag zu dessen Reduktion leisten könnte. Aber augenblicklich sollte man im Licht dieser Arbeit auf andere Möglichkeiten der Kalibrierung vertrauen, wenn man biasbehaftete Ausfallprozesse korrigieren möchte. Insbesondere die Anwendung eines GREGs oder verwandter Methoden ist sicher vorzuziehen.

Literatur

AAPOR (2008). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. Technical report, The American Association for Public Opinion Research.

Alho, J. M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika 3*, 617–624.

Allport, G. (1937). *Personality - A Psychological Interpretation*. New York: Holt, Rinehart, & Winston.

Allport, G., P. E. Vernon, and G. A. Lindzey (1970). *A study of values*. Boston: Houghton Mifflin.

Amelang, M. (2006). *Differentielle Psychologie Und Persönlichkeitsforschung* (6., vollst. überarb. Aufl ed.). Stuttgart: Kohlhammer.

Angleitner, A. and F. Ostendorf (2004). *NEO-PI-R NEO-Persoenlichkeitsinventar nach Costa und McCrae*. goettingen: Hogrefe.

Archer, T. M. (2008). Response Rates to Expect from Web-Based Surveys and What to Do About It. *Journal of Extension 46*(3), n.n.

Baily, R. (2008). *Design of Comparative Experiments*. Cambridge: Cambridge University Press.

Bamberg, S., E. Davidov, and P. Schmidt (2008). *Rational Choice: Theoretische Analysen und empirische Resultate. Festschrift für Karl-Dieter Opp zum 70. Geburtstag*, Chapter Wie gut erklären enge oder weite Rational-Choice Versionen Verhaltensänderungen?, pp. 143–171. VS Verlag für Spozialwissenschaften.

Bandilla, W., M. Bosnjak, and P. Altdorfer (2003). Survey Administration Effects?: A Comparison of Web-Based and Traditional Written Self-Administered Surveys Using the ISSP Environment Module. *Social Science Computer Review 21*(2), 235–249.

Bassili, J. (1993). Response latency versus certainty as indexes of the strength of voting intentions in a CATI survey. *Public Opinion Quarterly 57*, 54–61.

Batinic, B. (2002). *Online Social Sciences*. Hogrefe.

Benaglia, T., D. Chauveau, and D. R. Hunter (2009). Bandwidth Selection in an EM-like algorithm for nonparametric multivariate mixtures. *Journal of Computational and Graphical Statistics 18*, 505–526.

Benaglia, T., D. Chauveau, D. R. Hunter, and D. Young (2009). mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software 32*(6), 1–29.

Bethlehem, J. (1988). Reduction of the nonresponse bias through regression estimation. *Journal of official Statistics 4*, 251–260.

Bethlehem, J. (2002). *Survey Nonresponse*, Chapter Weighting Nonresponse Adjustments Based on Auxiliary Information, pp. 275–288. Hoboken, NJ: John Wiley & Sons.

Bethlehem, J. (2007). *Reducing the Bias of Web Survey Based Estimates*. Voorburg: Statistics Netherlands.

Bethlehem, J. (2009). *Applied Survey Methods*. Hoboken, NJ: John Wiley & Sons.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review 78*(2), 161–188.

Biemer, P. and L. Lyberg (2003). *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons.

Bizer, G. Y., P. S. Visser, M. K. Berent, and J. A. Krosnick (2004). *Studies in public opinion: Gauging attitudes, nonattitudes, measurement error and change*, Chapter Exploring the latent structure of strength-related attitude attributes, pp. 215–241. Princeton: Princeton University Press.

Bonke, J. and P. Fallesen (2010). The impact of incentives and interview methods on response quantity and quality in diary- and booklet-based surveys. *Survey Research Methods 4*(2), 91–101.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. ter Weel (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources 43*, 972–1059.

Borkenau, P. and F. Ostendorf (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen: Hogrefe.

Bosnjak, M. (2002). *(Non)Response bei Web-Befragungen*. Ph. D. thesis, University of Mannheim, Germany.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984). *Classification and Regression Trees*. London: Chapman and Hall.

Bundesamt für Statistik (2011). Internetnutzung in den Haushalten der Schweiz: Ergebnisse der Erhebung 2011 und Indikatoren. Technical report, BFS.

Bungard, W. (1980). *Die «gute» Versuchsperson Denkt Nicht: Artefakte in der Sozialpsychologie*. München: Urban & Schwarzenberg.

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality, 11*, 303–315.

Burnham, K. P. and D. R. Anderson (2004). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York et. al.: Springer.

Cacioppo, J. T. and R. E. Petty (1980). Sex differences in influenceability: Toward specifying the underlying processes. *Personality and Social Psychology Bulletin 6*, 651–656.

Callegaro, M. and C. Disogra (2008). Comouting Response Metrics for Online Panels. *Public Opinion Quarterly 72*(5), 1008–1032.

Camerer, C. (2003). *Behavioral game theory - experiments in strategic interaction*. Princton: Princton University Press.

Cassel, C. M., C.-E. Särndal, and J. H. Wretman (1983). *Incomplete Data in Sample Surveys, Vol. 3, Proceedings of the Symposium*, Chapter Some Use of Statistical Models in Connection with the Nonresponse Problem, pp. n.n. New York: Academic Press.

Chaudhuri, A. (2008). *Experiments in Economics: Playing Fair With Money*. New York: Routledge.

Church, A. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly 57*(3), 62–79.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association 368*(74), 829–836.

Cleveland, W. S. and S. J. Devlin (1988). Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association 403*(83), 596–610.

Cobben, F. (2009). *Nonresponse in Sample Surveys: Methods for Analysis and Adjustment*. The Hague: Statistics Netherlands.

Cobben, F. and J. Bethlehem (2005). Adjusting Undercoverage and Non-Response Bias in Telephone Surveys. Discussion Paper 05006, Statistics Netherlands, Voorburg/Heerlen.

Cochran, W. G. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics 24*, 295–313.

Cochran, W. G. (1977). *Sampling techiques*. Hoboken, NJ: John Wiley & Sons.

Collett, D. (2002). *Modelling Binary Data*. Taylor and Francis Ltd.

Comley, P. (2007). *Market Research Handbook*, Chapter Online Market Research., pp. 401–420. Hoboken, NJ: John Wiley & Sons.

Couper, M. P. (2007). Issues of Representation in Ehealth Research (with a Focus on Web Surveys). *American Journal of Preventive Medicine 35*(5S), 83–89.

Couper, M. P., R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls, and J. M. O'Reilly (1998). *Computer Assisted Survey Information Collection*. Hoboken, NJ: John Wiley & Sons.

Da Silva, D. N. and J. D. Opsomer (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *The Canadian Journal of Statistics 4*, 563–579.

Da Silva, D. N. and J. D. Opsomer (2008). Theoretical properties of propensity weighting for survey nonresponse through local polynomial regression. Technical Report Technical Report 2008/6, Department of Statistics, Colorado State University.

Da Silva, D. N. and J. D. Opsomer (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology 35*(2), 165–176.

D'Agostino, R. B. and D. B. Rubin (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association 95*(451), 749–759.

David, M., R. Little, M. Samuhel, and R. Triest (1983). Imputation models based on the propensity to respond. *ASA Proceedings of the Business and Economic Statistics Section N.N.*, 168–173.

Davidov, E. (2002). A Cross-Country and Cross-Time Comparison of the Human Values Measurements with the Second Round of the European Social Survey. *Survey Research Methods 2*(1), 33–46.

Davidov, E., P. Schmidt, and S. Schwartz (2008). Bringing Values Back In: The Adequacy of the European Social Survey to Measure Values in 20 Countries. *Public Opinion Quarterly 72*(3), 420–445.

Davidson, A. R., S. Yantis, M. Norwood, and D. E. Montano (1985). Amount of Information About Attitude Object and Attitude-Behavior Consistency. *Journal of Personality and Social Psychology 49*, 1184–1198.

de Heij, V., B. Schouten, and N. Shlomo (2010). *RISQ manual Tools in SAS and R for the computation of R-indicators and partial R-indicators., Work package 8, Deliverable 12.1*.

de Leeuw, E. (1992). *Data Quality in Mail, Telephone and Face-To-Face-Interviews*. Amsterdam: TT-Publikatjes.

de Vaus, D. (2002). *Social Surveys: Sage Benchmarks in Social Research Methods (4 Vols.)*. London: Sage Publications.

Dehne, M. and J. Schupp (2007). *Persönlichkeitsmerkmale im Sozio-oekonomischen Panel (SOEP) - Konzept, Umsetzung und empirische Eigenschaften*. Berlin: DIW Research Notes 26.

DeMaio, T. J. (1980). Refusals: Who, Where, and Why. *Public Opinion Quarterly 44*(2), 223–233.

DeMaio, T. J. (Ed.) (1983). *Statistical Policy Working Paper 10 - Approaches to Developing Questionnaires*. Bureau of the Census.

Deville, J. and C.-E. Särndal (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association 87*(418), 376–382.

Diekmann, A. (2007). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen* (18. Aufl. ed.). Rowohlt Tb.

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design* (2nd updated ed.). Hoboken, NJ: John Wiley & Sons.

Dillman, D. A., R. D. Tortora, J. Conradt, and D. Bowker (1998). Influence of Plain vs. Fancy Design on Response Rates for Web Surveys. In *Joint Statistical Meetings*, Dallas, Tex.

Downs, E., C. W. Smeyak, and M. G. Paul (1980). *Professional Interviewing*. Harpercollins College Div.

Drake, C. (1993). Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics 49*, 1231–1236.

Drewes, F. (2010). Is Data Impressed by Propensity Weighting – An Empirical Test. In *General Online Research Conference, Pforzheim*.

Ekholm, A. and S. Laaksonen (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. *Journal of Official Statistics 7*(3), 325–337.

El-Menouar, Y. and J. Blasius (2005). Abbrüche bei Online-Befragungen. Ergebnisse einer Befragung bei Medizinern. *ZA-Information 56*, 70–92.

Erbslöh, B. and A. Koch (1988). Die Non-Response-Studie zum ALLBUS 1986. Problemstellung, Design, erste Ergebnisse. *ZUMA Nachrichten 22*, 29–44.

ESOMAR (2005). Conducting Marcet and Opinion Research Using the Internet (Esomar World Research Codes and Guidelines). Technical report, European Society for Opinion and Marketing Research.

Esser, H. (1973). *Studien zum Interview*, Chapter Kooperation und Verweigerung beim Interview, pp. 69–142. Meisenheim am Glan: Verlag A. Hain.

Esser, H. (1986). Über die Teilnahme an Befragungen. *ZUMA Nachrichten 18*, 38–47.

Esser, H. (1996). What is Wrong with Variable Sociology? *European Sociological Review 12*(2), 159–166.

Ezzati-Ricea, T. M., M. R. Frankel, D. C. Hoaglinb, J. D. Loft, V. G. Coronadod, and R. A. Wright (2000). An alternative measure of response rate in random-digit-dialing surveys that screen for eligible subpopulations. *Journal of Economic and Social Measurement 26*, 99–109.

Fehr, E. and H. Gintis (2007). Human Motivation and Social Cooperation: Experimental and Analytical Foundations. *Annual Review of Sociology 33*(1), 43–64.

Feldman, S. (2003). *Oxford Handbook of Political Psychology*, Chapter Values, Ideology, and Structure of Political Attitudes, pp. 477–508. New York: Oxford University Press.

Fiske, A. P., S. Kitayama, H. R. Markus, and R. E. Niesbett (1998). *The Handbook of Social Psychology, 4th Edition*, Volume 2, Chapter The social matrix of social psychology, pp. 915–981. Boston: Oxford University Presas (McGraw Hill).

Folsom, R. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *ASA Proceedings of the Business and Economic Statistics Section N.N.*, 197–202.

Forsa (2006). Akzeptanz von Interviews. Telefonmarketing: Überwiegend als Belästigung erlebt. *Contex 08/06*, 2–6.

Fox, J. and S. Weisberg (2010). *car: Companion to Applied Regression*. R package version 2.0-2.

Fricker, R. D. and M. Schonlau (2002). Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature. *Field Methods 14*(4), 347–367.

Fricker, S., M. Galesic, R. Tourangeau, and T. Yan (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinion Quarterly 69*, 370–392.

Friedman, H. S. and M. W. Schustack (2005). *Personality: Classic Theories and Modern Research* (0003 ed.). Allyn & Bacon.

Froidevaux, Y. and V. Täube (2006). Internetnutzung in den Haushalten der Schweiz: Ergebnisse der Erhebung 2004 und Indikatoren. In

Statistik der Schweiz. Neuchâtel: Bundesamt für Statistik.

Gabler, S. and S. Häder (2007). Haushalts- und Personenerhebungen: Machbarkeit von Random Digit Dialing in der Schweiz. Technical report, Bundesamt für Statistik.

Galesic, M. (2002). Effects of questionnaire length on response rates: review of findings and guidelines for future research. In *German Online Research Conference (GOR), 2002*.

Galesic, M. (2005). Effects of questionnaire length on quality of responses in web surveys. In *ESF Workshop on Internet survey methodology, Dubrovnik, 2005*.

Ganassali, S. (2008). The Influence of the Design of Web Survey Questionnaires on the Quality of Responses. *Survey Research Methods 2*(1), 21–32.

Gelman, A., Y.-S. Su, M. Yajima, J. Hill, M. G. Pittau, J. Kerman, and T. Zheng (2010). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.3-05.

Gerlitz, J.-Y. and J. Schupp (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *Research Notes des Deutschen Instituts für Wirtschaftsforschung (DIW) Berlin 4*, 1–44.

Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princton: Princeton University Press.

Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron 4*, 184–200.

Göritz, A. (2007). *The Oxford Handbook of Internet Psychology*, Chapter Using Online Panels in Psychological Research, pp. 473–485. Norfolk: Oxford University Press.

Göritz, A. and H.-G. Wolf (2008). The long-term effect of material incentives on participation in online panels. *Behavior Research Methods 40*(4), 1144–1149.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality 37*, 504–528.

Goyder, J. (1987). *The Silent Minority. Nonrespondents on Sample Surveys*. Cambridge: Polity Press.

Groves, R. (1989). *Survey Error and Survey Costs*. Hoboken, NJ: John Wiley & Sons.

Groves, R., R. B. Cialdini, and M. P. Couper (1992). Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly 54*, 475–495.

Groves, R. and M. P. Couper (1998). *Nonresponse in Household Interview Surveys*. Hoboken, NJ: John Wiley & Sons.

Groves, R., D. Dillman, J. Eltinge, and R. Little (2002). *Survey Nonresponse*. Hoboken, NJ: John Wiley & Sons.

Groves, R., F. Fowler, M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons.

Groves, R. and E. Peytcheva (2008). The Impact of Nonresponse Rates on Nonresponse Bias. A Meta-Analysis. *Public Opinion Quarterly 72*(2), 167–189.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly 70*, 646–675.

Hartmann, P. and B. Schimpel-Neimanns (1992a). Zur Repräsentativität soziodemographischer Merkmale des Allbus – Multivariate Analysen zum Mittelschichtbias in der Umfrageforschung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie 44*(2), 315–340.

Hartmann, P. H. and B. Schimpel-Neimanns (1992b). Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage. *Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage 44*(2), 315–340.

Hartmann, P. H. and B. Schimpel-Neimanns (1993). Affirmative Repräsentativitätsbeweise oder Test konkreter Hypothesen zu Verteilungsabweichungen? *Kölner Zeitschrift für Soziologie und Sozialpsychologie 45*(2), 359–365.

Hastie, T., R. Tribshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York et. al.: Springer.

Heerwegh, D. (2009). Mode Difference between Face-To-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. *International Journal of Public Opinion Research 21*(1), 111–121.

Heerwegh, D. and G. Loosveldt (2002). An evaluation of the effect of response formats on data quality in web surveys. In *presented at the International Conference on Improving Surveys*, Copenhagen.

Hembroff, L. A., D. Rusz, A. Rafferty, H. McGee, and N. Ehrlich (2005). The Cost-Effectiveness of Alternative Advance Mailings in a Telephone Survey. *Public Opinion Quarterly 69*(2), 246–263.

Herzog, R. and J. Bachman (1981). Effects of Questionnaire Length on Response Quality. *Public Opinion Quarterly 45*(4), 549–559.

Hoogendoorn, A. W. and J. Daalmans (2009). Nonresponse in the Recruitment of an Internet Panel Based on Probability Sampling. *Survey Research Methods 3*(2), 59–72.

Horvitz, D. and D. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Huggins, V. and J. Eyerman (November 2001). Probability Based Internet Surveys: A Synopsis of Early Methods and Survey Research Results. In *Federal Committee on Statistical Methodology Research Conference*, Arlington.

Hulliger, B. (2006). Einführung in die Methoden der Stichprobenerhebung. Technical report, Bundesamt für Statistik, Neuchâtel.

Hulliger, B., A. Ries, T. Comment, and A. Bender (1997). *Proceedings of the Conference on Statistical Science honoring the bicentennial of Stefano Franscini's birth*, Chapter Weighting the Swiss Labour Force Survey, pp. 169–181. Basel: Birkhäuser.

John, O. P., E. M. Donahue, and R. L. Kentle (1991). *The Big Five Inventory-Versions 4a and 54*. Ph. D. thesis, University of California, Berkeley,

Institute of Personality and Social Research.

Kalfs, N. and W. Saris (1998). Large Differences in Time Use for Three Data Collection Systems. *Social Indicators Research 44*(3), 267–290.

Katz, J. and A. Tassone (1990). Public Opinion Trends: Privacy and Information Technology. *Public Opinion Quarterly 54*(1), 125–143.

Kim, J. and J. Kim (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics 4*, 501–514.

Kirchgässner, G. (2008). *Homo Oeconomicus: The Economic Model of Individual Behavior and Its Applications in Economics and Other Social Sciences*. New York et. al.: Springer.

Klein, D., C. A. Roster, G. Albaum, and R. Rogers (2004). A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research 46*(3), –.

Knapp, F. & Heidingsfelder, M. (2001). *Dimensions of Internet Science*, Chapter Ulf-Dietrich Reips and Michael Bosnjak. Lengerich: Pabst Science Publishers.

Kohm, M. and J.-U. Morawski (2008). *KOMA-Script Eine Samlung von Klassen und Paketen für LaTeX2e*. Lehmanns Media.

Kühne, M. and R. Böhme (2006). Effekte von Informationsstand, Wissen und Einstellungsstärke von Befragten auf die Antwortstabilität in Online-Befragungen mit Selbstrekrutierung. *ZUMA Nachrichten 59*, 42–72.

Kutner, M. H., C. Nachtsheim, J. Neter, and W. Li (2004). *Applied linear statistical models* (5th ed.). Open University Press, McGraw Hill Publ.Comp.

Lakatos, I. (1995). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.

Lamport, L. (1994). *LaTeX: A document preparation system: User's guide and reference*. Reading, Mass: Addison-Wesley Professional.

Lang, F. R., O. Lüdtke, and J. B. Asendorpf (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory

(BFI) bei jungen, mittelalten und alten Erwachsenen. *Diagnostica 47*(3), 111–121.

Larsen, R. J. and D. M. Buss (2007). *Personality Psychology: Domains of Knowledge About Human Nature* (3rd edition. ed.). Mcgraw Hill Publ.Comp.

Laux, L. and A. Gessner (2008). *Persönlichkeitspsychologie* (2. überarb. und erw. Aufl ed.). Stuttgart: Kohlhammer.

Leamer, E. E. (1983, March). Let's Take the Con Out of Econometrics. *American Economic Review 73*(1), 31–43.

Lee, S. (2006a). An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey. *Social Science Computer Review 24*(4), 460–475.

Lee, S. (2006b). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics 22*(2), 329–349.

Lin, I.-F. and N. C. Schaeffer (1995). Using Survey Participants to Estimate the Impact of Nonparticipation. *Public Opinion Quarterly 59*(2), 236–258.

Lipps, O. (2007). Cross-sectional and Longitudinal Interviewer and Respondent Effects in a CATI Panel Survey. In *ESRA Conference*, Prague.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove et.al.: Duxbury Press.

Lumley, T. (2010). *survey: R package version 3.22-1*.

Manfreda, K. L. and V. Vehovar (2002a). Mode Effects in Web Surveys. *American Association for Public Research 2002 N/A*, 2127–2177.

Manfreda, K. L. and V. Vehovar (2002b). Survey Design Features Influencing Response Rates in Web Surveys. In *The International Conference on Improving Surveys, Copenhagen*.

McCrae, R. R. and P. T. Costa (1985). *The NEO Personality Inventory - Manual*. Odessa: Psychological Assessment Resources.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Hoboken, NJ: John Wiley & Sons.

Meyers, S. and J. Oliver (1978). Privacy and Hostility Toward Government As Reasons for Refusal. *Proceedings of the Social statistics Section, American Statistical Association n. n.*, 509–513.

Morgan, J. and J. Sonquist (1963). Problems in the Analysis of Survey Data and a Proposal. *Journal of the American Statistical Association 58*, 415–434.

Nederhof, A. J. (1987). When neutrality is negative. *Quality and Quantity 21*(4), 425–432.

Neller, K. (2005). Kooperation und Verweigerung: eine Nonresponse Studie. *ZUMA Nachrichten 57*, 9–36.

Oberski, D. (2006). Correcting for bias in a self-selected sample using a random sample as auxiliary information: the Dutch Mobility Study. In *SMABS-EAM Conference 2006*.

Opp, K.-D. (1976). *Individualistische Sozialwissenschaft : Arbeitsweise Und Probleme Individualistisch Und Kollektivistisch Orientierter Sozialwissenschaften*. Stuttgart: Ferdinand Enke.

Pervin, L. A., D. Cervone, and O. P. John (2005). *Persönlichkeitstheorien* (5., vollst. überarb. und erw. Aufl ed.). München: E. Reinhardt.

Petrie, R. S., D. Moore, and D. A. Dillman (1997). Establishment Surveys: The Effect of Multi-Mode Sequence on Response Rate. *Proceedings of the Survey Research Methods Section n. n.*, 981–988.

Popper, K. (1935). *Logik der Forschung*. Berlin: Julius Springer Verlag.

Porst, R. (2008). *Fragebogen*. VS Verlag für Spozialwissenschaften.

Porter, S. R. and M. E. Whitcomb (2003). The Impact of Lottery Incentives on Student Survey Response Rates. *Research in Higher Education 44*(4), 389–407.

Postoaca, A. (2006). *The Anonymous Elect. Market Research through Online Access Panels*. New York et. al.: Springer.

Preisendörfer, P. (1999). *Umwelteinstellungen und Umweltverhalten in Deutschland*. Opladen: Leske + Budrich.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rammstedt, B. and O. P. John (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality 41*, 203–212.

Reinecke, J. (1991). *Interviewer- und Befragtenverhalten : Theoretische Ansätze und methodische Konzepte*. Studien zur Sozialwissenschaft ; Bd. 106. Opladen: Westdeutscher Verlag.

Reips, U.-D. (2007). *The Oxford Handbook of Internet Psychology*, Chapter The methodology of Internet-based experiments, pp. 373–390. Oxford University Press.

Reis, H. T. (2000). *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press.

Reuband, K.-H. and J. Blasius (2000). *Methoden in Telefonumfragen*, Chapter Situative Bedingungen des Interviews, Kooperationsverhalten und Sozialprofil konvertierter Verweigerer. Ein Vergleich von telefonischen und face-to-face Befragungen, pp. 129–170. Wiesbaden: Westdeutscher Verlag.

Rhall, T. and B. Fine (2008). The quest for on-line quality research. In *ESOMAR 2008, Singapore*.

Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Ripley, B. (2009). *tree: Classification and regression trees*. R package version 1.0-27.

Robbins, A., L. Lamb, and E. Hannah (2008). *Learning the vi and vim Editors*. O'Reilly Media.

Roberts, B. W., R. W. Robins, K. H. Trzesniewski, and A. Caspi (2004). *Handbook of the Life Course*, Chapter Personality Trait Development in Adulthood, pp. 579–595. New York: Kluwer Academic/Plenum Publishers.

Robins, R. W. and R. C. Fraley (2007). *Handbook of Research Methods in Personality Psychology* (1 ed.). Guilford Publications.

Robinson, J. P. (1999). Activity patterns of time-diary dropouts. *Society and Leisure (Loisir et Société) 21*, 551–554.

Rokeach, M. (1973). *The nature of human values*. New York: Free Press.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica 70*(1), 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association 79*, 516–524.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Roszkowski, M. J. and A. G. Bean (1990). Believe it or not! longer questionnaires have lower response rates. *Journal of Business and Psychology 4*(4), 495–509.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons.

Rubin, D. B. and N. Thomas (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics 52*(1), 249–264.

Samuelson, P. A. and W. D. Nordhaus (1948). *Economics: An Introductory Analysis*. McGraw Hill.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York et. al.: Springer. ISBN 978-0-387-75968-5.

Särndal, C.-E. (1980). A Two-Way Classification of regression Estimation Strategies in Probability Sampling. *The Canadian Journal of Statistics / Le Revue Canadienne de Statistique 8*(2), 165–177.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Survey with Nonresponse*. Hoboken, NJ: John Wiley & Sons.

Särndal, C.-E. and S. Lundström (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. *Journal of Official Statistics 24*(2), 167–191.

Schafer, J. L. and J. W. Graham (2002). Missing Data: Our View of the State of the Art. *Psychological Methods 7*(2), 147–177.

Scherpenzeel, A. (2008). Online interviews and data quality: A multitrait-multimethod study. In *MESS Workshop, Zeist*.

Scherpenzeel, A. and J. G. Bethlehem (2010). *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, Chapter How representative are online panels? Problems of coverage and selection and possible solutions, pp. in press. Taylor & Francis.

Schleifer, S. (1986). Trends in Attitudes Toward and Participation in Survey Research. *Public Opinion Quarterly 50*(1), 17–26.

Schnauber, A. and G. Daschmann (2008). States oder Traits? Was beeinflusst die Teilnahmebereitschaft an telefonischen Interviews? *Methoden - Daten - Analysen 2*(2), 97–123.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmass, Entwicklung und Ursachen*. Opladen: Leske+Budrich.

Schoen, H. and T. Faas (2005). When Methodology Interferes With Substance: The Difference of Attitudes Toward E-Campaigning and E-Voting in Online and Offline Surveys. *Social Science Computer Review 23*(3), 326–333.

Schonlau, M., M. Schonlau, A. van Soest, A. Kapteyn, M. P. Couper, and J. Winter (2004). Adjusting for selection bias in Web surveys using propensity scores: the case of the Health and Retirement Study. *ASA Section on Survey Research Methods Session 149: Statistical Approaches for Web Survey Data*, 4326–4333.

Schonlau, M., A. van Soest, and A. Kaptayn (2007). Are Webographic or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring? *Survey Research Methods 1*(3), 155–163.

Schonlau, M., K. Zappert, L. P. Simon, K. Sansted, S. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. Berry (2002). A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review 21*(X), 1–11.

Schouten, B., F. Cobben, and J. Bethlehem (2009). Indicators for the Representativiness of Survey Response. *Survey Methodology 35*(1), 101–113.

Schroepler, J.-P. and G. Wagner (2005). Characteristics and Impact of Faked Interviews in Surveys – An Analysis of Genuine Fakes in the raw Data of SOEP. *Allgemeines Statistisches Archiv 89*(1), 7–20.

Schwartz, S. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology 25*, 1–65.

Schwartz, S. H., G. Melech, A. Lehmann, S. Burgess, M. Harris, and V. Owens (2001). Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross Cultural Psychology 32*, 519–42.

Schwarz, N. and L. A. Vaughn (2002). *Heuristics and biases: the psychology of intuitive judgement*, Chapter The Availability Heuristics Revisited: Ease of Recall and Content of Recall as Distinct Sources of Information, pp. 103–119. Cambridge: Cambridge University Press.

Seipel, C. and S. Eifler (2003). *Soziologie der Kriminalität*, Chapter Gelegenheiten, Rational Choice und Selbstkontrolle. Zur Erklärung abweichenden Verhaltens in High-Cost und Low-Cost Situationen, pp. 288–316. Kölner Zeitschrift für Soziologie und Sozialwissenschaft.

Sharp, L. M. and J. Frankel (1983). Respondent Burden: A Test of Some Common Assumptions. *Public Opinion Quarterly 47*, 36–53.

Sheets, T., A. Radlinski, J. Kohne, and G. A. Brunner (1974). Deceived Respondents: Once Bitten, Twice Shy. *Public Opinion Quarterly 38*(2), 261–263.

Shlomo, N., C. Skinner, B. Schouten, J. Bethlehem, and L.-C. Zhang (2009). Statistical Properties of R-indicators. Technical report, RISQ deliverable 2.1.

Sidenvall, B., C. Fjellström, J. Andersson, K. Gustafsson, U. Nygren1, and M. Nydah (2002). Reasons among older Swedish women of not participating in a food survey. *European Journal of Clinical Nutrition 56*(7), 561–567.

Simon, W. (2006). *Persönlichkeitsmodelle Und Persönlichkeitstests 15 Persönlichkeitsmodelle Für Personalauswahl, Persönlichkeitsentwicklung, Training Und Coaching*. Offenbach: GABAL-Verlag.

Singer, E., N. Mathiowetz, and M. P. Couper (1993). The Impact of Privacy and Confidentiality Concerns on Survey Participation the Case of the 1990 U. S. Census. *Public Opinion Quarterly 57*(4), 465–482.

Singer, E., J. van Heoewyk, and M. P. Maher (2000). Experiments with Incentives in Telephone Surveys. *Public Opinion Quarterly 64*(2), 171–188.

Smith, T. W. (1984). Estimating Nonresponse Bias with Temporary Refusals. *Sociological Perspectives 27*(4), 473–489.

Smith, T. W. (1995). Trends in Non-Response Rates. *International Journal of Public Opinion Research 7*(2), 157–171.

Sobol, M. G. (1959). Panel Mortality and Panel Bias. *Journal of the American Statistical Association 54*(258), 52–68.

Spates, J. L. (1983). The Sociology of Values. *Annual Review of Sociology 9*, 27–49.

Speizer, H., R. Baker, and K. Schneider (2005). Survey Mode Effects; Comparison between Telephone and Web. *Paper presented at the 60th annual meeting of the American Association For Public Opinion Association (APPOR) N/A*, N/A.

Stocké, V. and B. Becker (2004). Determinanten und Konsequenzen der Umfrageeinstellung. Bewertungsdimensionen unterschiedlicher Umfragesponsoren und die Antwortbereitschaft der Befragten. *ZUMA Nachrichten 54*, 89–16.

Stocké, V. and B. Langfeldt (2003a). Umfrageeinstellung und Umfrageerfahrung. *ZUMA Nachrichten 27*, 55–88.

Stocké, V. and B. Langfeldt (2003b). Umfrageklima in Deutschland. *Context 14/03*, 6–8.

Stoop, I. (2005). *The hunt for the Last Respondent: Nonresponse in sample surveys*. Amsterdam: Aksant Academic Publishers.

Stoop, I., R. Jowell, and P. Mohler (2002). The European Social Survey1: One Survey in Two Dozen Countries. In *International Conference on Improving Surveys*, Kopenhagen.

Su, J., P. Shao, and J. Fang (2008). Effect of Incentives on Web-Based Surveys. *Tsinghua Science & Technology 13*(3), 344–347.

Taylor, H. (2000). Does Internet research work? Comparing on-line survey results with telephone surveys. *International Journal of Market Research 42*(1), 51–63.

Templ, M., A. Alfons, and A. Kowarik (2010). *VIM: Visualization and Imputation of Missing Values*. R package version 1.4.

Terhanian, G. (2000). How to Produce Credible, Trustworthy Information through Internet-Based Survey Research. In *Paper presented at the annual meeting of the American Association for the Public Opinion Research, Portland, OR*.

Therneau, T. M., B. Atkinson, and B. Ripley (2009). *rpart: Recursive Partitioning*. R package version 3.1-45.

Thomas, R., R. Lafond, S. Behnke, and J. Krosnick (2003). Can What We Don't Know (about Don't Know) Hurt Us?: Effects of Item Non-response. In *Meeting of the American Association for Public Opinion Research, Sheraton Music City, Nashville 2003*.

Tillé, Y. and A. Matei (2009). *sampling: R package version 2.3*.

Toepoel, V., M. Das, and A. van Soest (2009). Relating Quation Type to Panel Condition: Comparing trained and fresh Respondents. *Survey Research Methods 3*(2), 73–80.

Tomaskovic-Devey, D., J. Leiter, and S. Thompson (1994). Organizational survey nonresponse. *Administrative Science Quarterly 39*(Sep.), n.n.

Tortora, R. (2009). *Methodology of Longitudinal Surveys*, Chapter Attrition in Consumer Panels, pp. 235–249. Horbroken, NJ: John Wiley & Sons.

Tourangeau, R. (2003). *Survey Automation. Report and Workshop Proceedings*, Chapter Web-Based Data Collection, pp. 183–193. National Academies Press.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000). *Psychology of Survey Response*. Cambridge Univ Press.

Tourangeau, R. and T. W. Smith (1998). *Computer Assisted Survey Information Collection*, Chapter Collecting sensitive information with different modes of data collection, pp. 431–454. Hoboken, NJ: John Wiley & Sons.

Tupes, E. C. and R. E. Christal (1961/1992). Recurrent Personality Factors Based on Trait Ratings. *Journal of Personality 60*(2), 225–251. Erstmals veröffentlicht in: Tech. Rep. No. ASD-TR-61-97, Lackland Air Force Base, US Air Force.

Urbina, S. (2004). *Essentials of Psychological Testing*. Hoboken, NJ: John Wiley & Sons.

van der Zouwen, J., S. Draisma, M. Eikelenboom, and J. Smit (2009). Analyzing Interviewer Bahaviour in Semi-Standardized Face-to-Face Interviews. In *ESRA conference*, Warsaw.

Varedian, M. and G. Forsman (2002). Comparing Propensity Score Weighting with Other Weighting Methods: A Case Study on Web Data. In *Paper presented at the American Association for Public Opinion Research Conference, St.Petersburg*.

Vehovar, V., Z. Batagelj, K. L. Manfreda, and M. Zaletel (2002). *Survey Nonresponse*, Chapter Nonresponse in Web Surveys, pp. 229–242. Hoboken, NJ: John Wiley & Sons.

Vehovar, V., K. L. Manfreda, and Z. Batagelj (1999). Web Surveys: Can The Weighting Solve The Problem? In *Paper presented at the American Association for Public Opinion Research Conference, St. Petersburg*.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York et. al.: Springer. ISBN 0-387-95457-0.

Verplanken, B. and R. W. Holland (2002). Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology 82*, 434–447.

Viseu, A., A. Clement, and J. Aspinall (2004). Situating privacy online: Complex perceptions and everyday practices. *Information, Communication and Society 7*(1), 92–114.

Visser, P. (1998). *Assessing the structure and function of attitude strength: Insights from a new approach*. Ph. D. thesis, The Ohio State University.

VSMS (2009). *Jahrbuch 2009*. Verband Schweizer Markt- und Sozialforscher.

VSMS (2010). *Jahrbuch 2010*. Verband Schweizer Markt- und Sozialforscher.

Wand, M. and B. Ripley (2010). *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*. R package version 2.23-4.

Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing*. Chapman & Hall (CRC Monographs on Statistics & Applied Probability).

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association 60*(309), 63–69.

Warriner, K., J. Goyder, H. Gjertsen, P. Hohner, and K. McSpurren (1996). Charities, no; Lotteries, no; Cash, yes: Main Effects and Interactions in a Canadian Incentives Experiment. *Public Opinion Quarterly 60*(2), 542–562.

Weil, F. (2005). *Tugenden der Medienkultur. Zu Sinn und Sinnverlust tugendhaften Handelns in der medialen Kommunikation*, Chapter Privatsphäre - schützenswert oder uncool?, pp. 107–119. Franz Steiner Verlag.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York et. al.: Springer.

Williams, P., R. van Ossenbruggen, and T. Vonk (2006). The Effects of Panel Recruitment and Management on Research Result. In *ESOMAR Panel Research*, Barcelona.

Wolf, C. (1995). Sozio-Ökonomischer Status und berufliches Prestige. *ZUMA Nachrichten 37*, 102–136.

Wood, W. (1982). Retrieval of attitude-relevant information from memory: Effects on susceptibility to persuasion and on intrinsic motivation. *Journal of Personality and Social Psychology 42*, 798–810.

Yun, G. W. and C. W. Trumbo (2000). Comparative Response to Survey Executed by Post, E-Mail and Web-Form. *Journal of Computer-Mediated Communication 6*(1), N/A.

Zeileis, A. and T. Hothorn (2002). Diagnostic Checking in Regression Relationships. *R News 2*(3), 7–10.

Abbildungsverzeichnis

Tabellenverzeichnis

A Abkürzungsverzeichnis

Die folgende Liste enthält die wichtigsten Abkürzungen in alphabetischen Reihenfolge[1].

CAI	Computer-assisted interviewing
CAPI	Computer-assisted personal interviewing
CASAQ	Computer-administered self-administered questionnaires
CASI	Computer-assisted self interviewing
CATI	Computer-assisted telephone interviewing
CAWI	Computer-assisted Web interviewing
ENK	Engerer Nutzerkreis
ESOMAR	European Society for Opinion and Marketing Research
ESRA	European Survey Research Asscociation
GREG	Generalized Regression Estimator
HI	Harris Interactive
IPF	Iterational Proportional Fiiting
LOESS	locally weighted scatterplot smoothing
MAR	Missing at Random
MCAR	Missing Completely at Random
NMAR	Not Missing at Random
PSA	Propensity Score Adjustment
RDD	Random Digit Dialing
RCT	Rational Choice Theory, Theorie rationaler Wahl
ρ (Rho)	Teilnahmewahrscheinlichkeit, Propensity Score
VSMS	Verband Schweizer Markt- und Sozialforscher
WNK	Weiterer Nutzerkreis
X	Erklärende variable, Kovariate
Y	Interessierende Variable

1 Die Abkürzungen werden alle ausführlich im Text eingeführt. Hier sind nur solche Abkürzungen aufgeführt, die wiederholt an verschiedenen Stellen des Textes verwendet werden. Solche, die nur einmalig vorkommen, wie insbesondere in Formeln oder bei den nur am Rande vorgestellten Varianeten des Big Five Inventories wie z. B. dem NEO-PI-RR, werden hier nicht aufgeführt. Auch die Variablennamen der Befragungen sind hier nicht nochmals erwähnt, sondern finden sich auf S. 95

B Auswahl der Items aus der Schwartz-Skala

Die Schwartz-Skala wird in ihrer aktuellen Version im European Social Survey auf folgende Art verwendet. Vor der jeweiligen Frage steht jeweils der Variablenname. Hier sind nicht alle Items aufgelistet, sondern nur diejenigen, die es in einer Vorauswahl auf grund ihrer Plausibilität ausgewählt wurden.

Alle Aussagen sind jeweils projektiv formuliert, d. h. die Befragten sollen verschiedene Eigenschaften be einer fiktiven Person mit sich vergleichen. Das Geschlecht der fiktiven Person wird im mündlichen Interview jeweils dem Geschlecht der Befragten angepasst.

Im folgenden beschreibe ich kurz einige Personen. Hören Sie den Beschreibungen aufmerksam zu. Entscheiden Sie jedes mal, ob Ihnen die Person sehr ähnlich, ähnlich, etwas ähnlich, nur ein kleines bisschen ähnlich, nicht ähnlich oder überhaupt nicht ähnlich ist. Wie ähnlich ist Ihnen diese Person?

ipcrtiv *Es ist ihr wichtig, neue Ideen zu entwickeln und kreativ zu sein. Sie macht Sachen gerne auf ihre eigene originelle Art und Weise.*

imprich *Es ist wichtig für sie, reich zu sein. Sie möchte viel Geld haben und teure Sachen besitzen.*

ipquopt *Sie hält es für wichtig, dass alle Menschen auf der Welt gleich behandelt werden sollten. Sie glaubt, dass jeder Mensch im Leben gleiche Chancen haben sollte.*

ipshabt *Es ist ihr wichtig, ihre Fähigkeiten zu zeigen. Sie möchte, dass die Leute bewundern, was sie tut.*

impsafe *Es ist ihr wichtig, in einer sicheren Umgebung zu leben. Sie vermeidet alles, was ihre Sicherheit gefährden könnte.*

impdiff *Sie liebt Überraschungen und hält immer Ausschau nach neuen Aktivitäten. Sie denkt, dass im Leben Abwechslung wichtig ist.*

impfrule *Sie glaubt, dass die Menschen tun sollten, was man Ihnen sagt. Sie denkt, dass Menschen sich immer an Regeln halten sollten, selbst dann, wenn es niemand sieht.*

ipudrst *Es ist ihr wichtig, Menschen zuzuhören, die anders sind als sie. Auch wenn sie anderer Meinung ist als andere, will sie sie trotzdem verstehen.*

ipmodest *Es ist ihr wichtig, zurückhaltend und bescheiden zu sein. Sie versucht, die Aufmerksamkeit nicht auf sich zu lenken.*

ipgtim *Es ist ihr wichtig, Spaß zu haben. Sie gönnt sich selbst gerne etwas.*

impfree *Es ist ihr wichtig, selbst zu entscheiden, was sie tut. Sie ist gerne frei und unabhängig von anderen.*

iphlppl *Es ist ihr sehr wichtig, den Menschen um sie herum zu helfen. Sie will für deren Wohl sorgen.*

ipsuces *Es ist ihr wichtig, sehr erfolgreich zu sein. Sie hofft, dass andere ihre Leistungen anerkennen.*

ipstrgv *Es ist ihr wichtig, dass der Staat ihre persönliche Sicherheit vor allen Bedrohungen gewährleistet. Sie will einen starken Staat, der seine Bürger verteidigt.*

ipadvnt *Sie sucht nach Abenteuern und nimmt gerne Risiken auf sich. Sie möchte ein aufregendes Leben führen.*

ipbhprp *Es ist ihr wichtig, sich immer richtig zu verhalten. Sie möchte vermeiden irgendetwas zu tun, von dem die Leute sagen könnten, dass es falsch ist.*

iprspot *Es ist ihr wichtig, von anderen respektiert zu werden. Sie will, dass die Leute tun, was sie sagt.*

iplylfr *Es ist ihr wichtig, gegenüber ihren Freunden loyal zu sein. Sie will sich für Menschen einsetzen, die ihr nahe stehen.*

impenv *Sie ist fest davon überzeugt, dass die Menschen sich um die Natur kümmern sollten. Umweltschutz ist ihr wichtig.*

imptrad *Tradition ist ihr wichtig. Sie versucht, sich an die Sitten und Gebräuche zu halten, die ihr von ihrer Religion oder Familie überliefert wurden.*

impfun *Sie lässt keine Gelegenheit aus, Spaß zu haben. Es ist ihr wichtig, Dinge zu tun, die ihr Vergnügen bereiten.*

C Der komplette Fragebogen

Der Fragebogen des Rekrutierungsinterviews teilt sich n sechs Teile. Obwohl es stilistisch nicht sehr schön ist, Fragen die zusammengehören nicht auch zusammen zu stellen (Porst, 2008), war dies notwendig. Zum Beispiel wirkt erfahrungsgemäss die Frage nach dem Einkommen abschreckend und wurde daher erst am Ende der Befragung gestellt, um die Zusage zur Panelteilnahme nicht zu gefährden[1].

In eckigen Klammern stehen jeweils die zugehörigen Fragen. Die einleitende Frage muss nur für die Auswahl der zu befragenden Person im Haushalt im Rahmen des Random-Quota-Auswahlverfahrens gestellt werden.

1. Demografische Angaben [2, 44 bis 50]
2. Kontrollfragen [4, 5, 6]
3. Fragen des Big Five TIPI Inventars [7 bis 21]
4. Fragen der Schwartz Skala [22 bis 31]
5. Fragen, die aus der Theorie Rationaler Handlungen abgeleitet wurden [3, 32 bis 38, 43]
6. Rekrutierungsfragen [39 bis 42]

Im folgenden ist der Fragebogen mit wichtigen Anweisungen an die Telefonistinnen wiedergegeben[2]. Diese und die Anweisungen an die EDV (Filtersetzung etc.) erscheinen in kursiver Schrift.

1 Die Änderung der Reihenfolge oblag den CATI-Spezialisten von LINK.

2 Anweisungen, wie z. B. in welchem Format eine Eingabe zu erfolgen hat, wurden weggelassen.

C.1 Deutsche Version

Grüezi, Link Forschungsinstitut, [*INT: Name*]. Wir machen zur Zeit eine Umfrage über aktuelle Themen. Dazu möchte ich Ihnen ein paar Fragen stellen. Doch zuerst, ist eine Person unter 24 Jahren in Ihrem Haushalt?

[*INT: Falls jemand sagt, es handle sich um eine geheime Nummer und wie wir zu dieser Nummer gekommen sind, antworten: Nummer wird vom Computer erstellt.*]

1. Darf ich fragen, wie alt Sie sind?
2. *Geschlecht des Befragten eingeben.*
3. Sind Sie voll, teilweise oder nicht erwerbstätig?
4. Welches ist Ihr Haupteinkaufsort für Lebensmittel und Artikel vom täglichen Bedarf? [*INT: Ohne Nennung zu kodieren als Coop, Migros, Denner und andere.*]
5. Besitzen Sie persönlich eines oder mehrere von den folgenden Abonnementen für den öffentlichen Verkehr?
   - Halbtax-Abo
   - Generalabonnement (GA)
   - Verbund-Abonnement für Ihre Region
   - Strecken-Abonnement (Monats- oder Jahres-Streckenabonnement)
   - Mehrfahrtenkarte
   - Gleis 7
   - Anderes Abonnement
   - Ich besitze keines dieser Abonnements.
6. Seit wann wohnen Sie an Ihrem jetzigen Wohnort?

   - seit diesem Jahr, also seit 2009
   - seit letztem Jahr, also 2008
   - seit ca. 2 Jahren (seit 2007)

- seit ca. 3 Jahren (seit 2006)
- schon seit mehr als 3 Jahren (seit 2005 oder früher)
- weiss nicht, k. A.

Die folgenden Aussagen beschreiben Eigenschaften und Einstellungen, wo auf einen mehr oder weniger zutreffen können. Bitte sagen Sie, inwiefern diese Aussagen Ihrer Meinung nach auf Sie zutreffen.

Sie können Ihre Antwort jeweils zwischen 1 und 5 abstufen. Dabei bedeutet 1 «trifft auf mich überhaupt nicht zu» und 5 bedeutet «trifft auf mich voll und ganz zu».

[*EDV: Items at random*]

[*INT: Vorlesen! Jeweils mit «Ich bin jemand, der. . .» beginnen*]

7. . . .gründlich schafft.
8. . . .kommunikativ und gesprächig ist
9. . . .manchmal ä chli grob zu anderen ist
10. . . .originell ist und Ideen einbringt
11. . . .sich oft Sorgen macht
12. . . .verzeihen kann
13. . . .eher faul ist
14. . . .aus sich herausgehen kann und gesellig ist
15. . . .künstlerische Erfahrungen schätzt
16. . . .leicht nervös wird
17. . . .Aufgaben wirksam und effizient erledigt
18. . . .zurückhaltend ist
19. . . .rücksichtsvoll und freundlich mit anderen umgeht
20. . . .eine lebhafte Phantasie hat
21. . . .entspannt ist und mit Stress gut umgehen kann

Auch die folgenden Aussagen beschreiben Eigenschaften und Einstellungen, wo auf einen mehr oder weniger zutreffen können. Bitte sagen Sie, inwiefern diese Aussagen Ihrer Meinung nach auf Sie zutreffen.

Sie können Ihre Antwort jeweils zwischen 1 und 5 abstufen. Dabei bedeutet 1 «trifft auf mich überhaupt nicht zu» und 5 bedeutet «trifft auf mich voll und ganz zu».

[*EDV: Items at random*]

[*INT: Vorlesen! Jeweils mit «Ich bin jemand, dem es wichtig ist,. . .» beginnen*]

22. . . .seine Fähigkeiten zu zeigen.

23. . . .in einer sicheren Umgebung zu leben.

24. . . .immer wieder neue Aktivitäten auszuprobieren.

25. . . .dem im Leben Abwechslung wichtig ist.

26. . . .sich immer an Regeln zu halten, selbst dann, wenn es niemand sieht.

27. . . .zurückhaltend und bescheiden zu sein.

28. . . .Spass zu haben.

29. . . .den Menschen um mir herum zu helfen.

30. . . .sich immer richtig zu verhalten und nicht den Eindruck zu erwecken, mich falsch zu verhalten.

31. . . .von anderen respektiert zu werden.

Folgende Aussagen beschreiben Ansichten und Meinungen zu Umfragen. Bitte sagen Sie uns jeweils, inwiefern Sie persönlich diesen Aussagen zustimmen.

Sie können Ihre Antwort jeweils zwischen 1 und 5 abstufen. Dabei bedeutet 1 «stimme überhaupt nicht zu» und 5 bedeutet «stimme voll und ganz zu».

32. Marktforschungsunternehmen behandeln die Daten vertraulich.

33. Umfragen bringen Abwechslung und sind interessant.

34. Bei Umfragen wird häufig öppes gefragt, was niemand etwas angeht.

35. Ich bin bereit, über meine persönlichen Gewohnheiten auch mit jemandem zu sprechen, den ich nicht so gut kenne.

36. Marktforschung ist für die Gesellschaft wichtig und sinnvoll.

37. Denken Sie jetzt einmal an die Nutzung vom Internet für private Zwecke, es spielt dabei keine Rolle, wo, d.h. von welchem Anschluss aus Sie das Internet nutzen. Wie häufig etwa nutzen Sie das Internet für private Zwecke? Ist das. . .
    - mehrmals täglich
    - einmal täglich
    - mehrmals pro Woche
    - einmal pro Woche
    - seltener
    - oder nie?
    - weiss nicht

    [*EDV: Filter setzen. Nächsten Fragen nur für Internetnutzer.*]

38. Wie oft haben Sie in den letzten 12 Monaten etwas online im Internet gekauft?

39. Haben Sie zu Hause einen Breitband-Internet-Anschluss mit ADSL oder Kabelmodem oder nicht? Bei einem Breitbandanschluss zahlen Sie eine fixe Monatsgebühr pro Monat für Ihren Internet-Anschluss. Sie können immer online sein, ohne Zusatzkosten zu haben.

40. Schaffen Sie oder öpper in Ihrem Haushalt in einer von den folgenden Branchen?
    - Marktforschung
    - Werbung oder PR
    - Journalist in Presse, Radio, Fernsehen oder Internet

41. Das LINK Institut macht auch Umfragen via Internet. Wir suchen Leute, wo bereit sind, regelmässig bei solchen Umfragen mitzumachen. Es gibt circa 1 Umfrage alle 2 Monate. Wir versichern Ihnen, dass

wir Ihnen nichts verkaufen werden, wir machen nur Forschung. Wir geben auch keine persönlichen Daten von Ihnen an Dritte weiter. Es bleiben auch keine Daten auf Ihrem PC gespeichert. Sie werden für jede Befragung eine kleine Entschädigung erhalten.

Sie können bei jeder Befragung auswählen zwischen Coop Superpunkten, Migros Cumuluspunkten, einem elektronischen Telefongutschein oder einer Spende für einen guten Zweck

[*INT: wechselnde Hilfswerke*]

Haben Sie Interesse, regelmässig bei dieser Form von Marktforschung mitzumachen?

- ja
- nein
- mache bereits beim LINK Panel mit

42. Wir schicken Ihnen in den nächsten 3 Wochen per Email einen Link zum Ausfüllen vom ersten Fragebogen. Können Sie mir bitte zu diesem Zweck Ihre Email-Adresse angeben?

43. Bitte sagen Sie uns jetzt noch, wie stark Sie persönlich der folgenden Aussage zustimmen. Sie können Ihre Antwort jeweils zwischen 1 und 5 abstufen. Dabei bedeutet 1 «stimme überhaupt nicht zu» und 5 bedeutet «stimme voll und ganz zu». Dass es für die Teilnahme an Online-Befragungen eine Belohnung gibt, ist für mich ein wichtiger Grund zum Mitmachen.

44. Ich möchte noch Ihren Namen und Vornamen aufnehmen.

45. Was ist Ihr Zivilstand?

46. Und wie setzt sich Ihr Haushalt zusammen? Wie viele Personen in Ihrem Haushalt sind:
    - Kinder bis 5 Jahre
    - Kinder 6- 9 Jahre
    - Kinder 10-14 Jahre
    - Jugendliche 15-19 Jahre

- Erwachsene 20-64 Jahre
- Erwachsene ab 65 Jahre

In Ihrem Haushalt wohnen also *X* Personen?

47. Welche Schule haben Sie zuletzt besucht?
    - keine Schule abgeschlossen
    - Primarschule
    - Sekundarschule
    - Berufsschule
    - Mittelschule/Seminar
    - Technikum/HWV/Fachhochschule
    - Universität/ETH/Hochschule

48. Damit wir diese Umfrage statistisch besser auswerten können, ist es für uns wichtig, dass wir noch Ihr Einkommen notieren können. Sie können sicher sein, dass auch diese Angabe streng vertraulich und anonym behandelt wird. Ich lese Ihnen zu diesem Zweck verschiedene Einkommensgruppen vor. Sagen Sie mir bitte, welches das Einkommen von allen Personen in Ihrem Haushalt zusammen ist?
    - bis Fr. 4'000
    - zwischen Fr. 4'001 und Fr. 6'000
    - zwischen 6'001 und Fr. 8'000
    - zwischen 8'001 und Fr. 10'000
    - zwischen 10'001 und Fr. 15'000
    - Mehr als Fr. 15'000
    - keine Angabe

49. Und welches monatliche Brutto-Einkommen trifft auf Sie persönlich zu? Auch diese Antwort wird natürlich streng vertraulich behandelt.
    - bis Fr. 4'000
    - zwischen Fr. 4'001 und Fr. 6'000

- zwischen 6'001 und Fr. 8'000
- zwischen 8'001 und Fr. 10'000
- zwischen 10'001 und Fr. 15'000
- Mehr als Fr. 15'000
- keine Angabe

50. Wie heisst die Postleitzahl von Ihrem Wohnort?

Damit sind wir am Schluss vom Interview. Herzlichen Dank. Es kann vorkommen, dass wir für eine Nachfrage oder bei einer Unklarheit nochmals kurz anrufen müssten. Das passiert allerdings selten. Wir wünschen Ihnen also noch einen schönen Abend und bedanken uns bei Ihnen für Ihre wertvollen Auskünfte.

Zusätzlich werden als Paradaten der Anfang und das Ende des Interviews protokolliert.

C.2 Französische Version

Bonsoir, ici ... [*ENQ.: INDIQUER VOTRE NOM*], de l'Institut de recherche LINK de Lausanne. Nous réalisons actuellement une enquête sur des sujets actuels. J'aimerais vous poser quelques questions sur ces sujets. Tout d'abord, y a-t-il dans votre ménage une personne de moins de 24 ans? [ENQ: SI QUELQU'UN DÉCLARE QU'IL S'AGIT LÀ D'UN NUMÉRO CONFIDENTIEL ET DEMANDE COMMENT NOUS L'AVONS OBTENU, LUI RÉPONDRE: CE NUMÉRO A ÉTÉ ÉTABLI / GÉNÉRÉ PAR ORDINATEUR.]

1. Tout d'abord puis-je vous demander votre âge ?
2. ENQ.: INDIQUER LE SEXE DE LA PERSONNE INTERROGEE
3. Exercez-vous une activité professionnelle à plein temps, à temps partiel ou n'exercez-vous pas d'activité professionnelle ?
4. Dans quel(s) magasin(s) ou supermarché(s) faites vous généralement vos achats alimentaires et autres articles pour vos besoins quotidiens? Migros, Coop, Denner ou autres

5. Possédez-vous personnellement un ou plusieurs des abonnements suivants pour les transports publics ?
   - Abonnement demi-tarif
   - Abonnement général (AG)
   - Abonnement lié à votre région
   - Abonnement de parcours (abonnement de parcours mensuel ou annuel)
   - Carte multi-course
   - Voie 7
   - Autre abonnement
   - Je ne possède aucun de ces abonnements
6. Quand avez-vous aménagé dans votre domicile? [INT: NE PAS LIRE; DEMANDER POUR CONFIRMER L'ANNÉE]
   - depuis cette année, donc depuis 2009
   - depuis un an, donc 2008
   - depuis environ 2 ans (ENQ: depuis 2007)
   - depuis envion 3 ans (ENQ: depuis 2006)
   - déjà depuis plus que 3 ans (ENQ: depuis 2005 ou avant)
   - ne sais pas/ pas de réponse

Les énonciations suivantes décrivent des qualités et points de vue qui peuvent s'appliquer plus ou moins sur une personne. S'il vous plaît, répétez, dans quelle mesure ces énonciations s'appliquent à vous.

Vous pouvez à chaque fois graduer votre réponse entre 1 et 5. 1 signifie «ne s'applique pas du tout à moi» et 5 signifie «s'applique tout à fait à moi». [*EDV: Items at random*]

[LIRE! En commençant avec «Je suis une personne qui»]

7. . . .travaille avec précaution.
8. . . .est communicative et bavarde
9. est parfois un peu grossière envers d'autres personnes

10. est originale et apporte des idées

11. se fait souvent des soucis

12. sait pardonner

13. est plutôt paresseuse

14. est sociable et qui peut s'éclater (s'amuser)

15. apprécie des expériences artistiques

16. s'énerve facilement

17. accomplit des tâches efficacement

18. est discrète

19. traite d'autres personnes avec plein d'égard et avec gentillesse

20. a une imagination vive

21. est relaxe et qui sait gérer le stress

Les énonciations suivantes décrivent des qualités et points de vue qui peuvent s'appliquer plus ou moins sur une personne. S'il vous plaît, répétez, dans quelle mesure ces énonciations s'appliquent à vous.

Vous pouvez à chaque fois graduer votre réponse entre 1 et 5. 1 signifie «ne s'applique pas du tout à moi» et 5 signifie «s'applique tout à fait à moi». [*EDV: Items at random*]

[Lire!] Je suis une personne pour qui, il est important de

22. . . .montrer ces capacités.

23. . . .vivre dans un environnement sécurisé

24. . . .toujours expérimenter de nouvelles activités

25. . . .d'avoir une vie variée.

26. . . .toujours respecter les règles, même si personne ne le remarque.

27. . . .être réticente et modeste.

28. . . .avoir du plaisir.

29. . . .aider les gens autour de moi.

30. . . .toujours se comporter comme il le faut et de ne pas donner l'impression de se comporter de manière incorrecte.

31. . . .être respecté par les autres.

Les énonciations suivantes décrivent des points de vues et des opinions pour des sondages. S'il vous plait, dites-nous dans quelle mesure approuvez-vous personnellement ces énonciations.

Vous pouvez à chaque fois graduer votre réponse entre 1 et 5. 1 signifie «j'approuve pas du tout» et 5 signifie «j'approuve tout à fait».

32. Des instituts de sondage traitent confidentiellement les données.

33. Des sondages amènent (apportent) la variation et sont intéressants.

34. Lors des sondages, on pose souvent des questions qui concernent la sphère privée.

35. Je suis disposé à parler de mes habitudes personnelles avec quelqu'un que je ne connais pas si bien.

36. L'étude de marché est importante et judicieuse pour la société.

37. Pensez maintenant à l'utilisation d'Internet à des fins privées. Le raccordement à partir duquel vous utilisez Internet ne joue aucun rôle ici. A quelle fréquence utilisez-vous l'Internet approximativement à des fins privés? Es-ce. . .
    - plusieurs fois par jour
    - une fois par jour
    - plusieurs fois par semaine
    - une fois par semaine
    - plus rarement
    - ou jamais?
    - ne sait pas

38. Au cours des 12 derniers mois, combien de fois avez-vous acheté quelque chose online sur Internet?

39. Avez-vous à votre domicile un accès Internet à large bande avec ADSL ou modem câblé? Pour un accès à large bande, payez-vous une taxe fixe mensuelle pour votre raccordement à Internet, et disposez-vous sans frais supplémentaires du raccordement grâce auquel vous êtes en permanence online?

40. Travaillez-vous vous-même ou quelqu'un d'autre de votre ménage travaille-t-il dans l'un des secteurs d'activités suivants?

    - Etudes de marché
    - Publicité ou relations publiques
    - Journalisme de presse, radio, télévision ou Internet

41. Notre Institut LINK réalise des enquêtes par Internet également. Nous recherchons des personnes prêtes à participer régulièrement à de telles enquêtes. Nous aimerions que vous remplissiez environ 1 questionnaire tous les 2 mois.

    Nous vous assurons que nous ne chercherons pas à vous vendre quoi que ce soit, car nous effectuons uniquement des recherches. Nous ne transmettons pas non plus aucune donnée personnelle à des tiers. Aucune donnée ne restera chargée dans votre PC.

    Pour chaque enquête, vous recevrez une petite indemnité. Vous pourrez à chaque fois choisir entre des superpoints Coop, des points Migros Cumulus, un bon électronique de téléphone ou un don dans un but caritatif

    [ENQ.: ouvres de charité diverses]

    Voyez-vous un intérêt à participer régulièrement à cette forme d'études de marchés ?

    - Oui
    - Non
    - participe déjà au panel internet link

42. Dans les 3 semaines environ, nous vous ferons parvenir un lien pour remplir le premier questionnaire sur Internet. Pour cela, j'aimerais bien maintenant relever votre adresse Email.

43. S'il vous plait dites-nous encore, dans quelle mesure approuvez-vous personnellement l'énoncé suivant:

    Vous pouvez graduer votre réponse entre 1 et 5. 1 signifie «j'approuve pas du tout» et 5 signifie «j'approuve tout à fait».

    Une raison importante pour la participation à un sondage en ligne est de recevoir une récompense.

44. J'aimerais encore contrôler votre nom et prénom.

45. Quel est votre état civil ?

46. Comment se compose votre ménage? Combien de personnes de votre ménage sont:

    - des enfants de 0 a 5 ans
    - des enfants de 6 a 9 ans
    - des enfants de 10 a 14 ans
    - des adolescents de 15 a 19 ans
    - des adultes de 20 a 64 ans
    - des adultes de 65 ans et plus

    Dans votre ménage vivent alors X personnes ?

47. Quelle école avez-vous fréquentée en dernier lieu ?

    - n'a pas terminé une école
    - école primaire
    - collège, école secondaire
    - école professionnelle
    - gymnase (bac/matu)
    - école technique / ESCEA / HES
    - université / EPFL

- pas de réponse

48. Pour que nous exploitions au mieux cette enquête du point de vue statistique, il est important pour nous que nous notions encore votre revenu. Vous pouvez d'ores et déjà être rassuré(e) que ces indications seront traitées confidentiellement et anonymement. A cet effet, je vous lis les catégories de revenus différentes Dites - moi, s'il vous plaît, lequel est le revenu total de toutes les personnes faisant parties dans votre budget(ménage)?
    - plus petit que Fr. 4'000.–
    - de Fr. 4'0001.– a Fr. 6'000.–
    - de Fr. 6'0001.– a Fr. 8'000.–
    - de Fr. 8'0001.– a Fr. 10'000.–
    - de Fr. 10'0001.– a Fr. 15'000.–
    - de plus de Fr. 15'000.–
    - pas de réponse
49. Et quel revenu mensuel brut s'applique personnellement à vous ? Aussi cette réponse sera évidemment traitée confidentiellement.
    - plus petit que Fr. 4'000.–
    - de Fr. 4'0001.– a Fr. 6'000.–
    - de Fr. 6'0001.– a Fr. 8'000.–
    - de Fr. 8'0001.– a Fr. 10'000.–
    - de Fr. 10'0001.– a Fr. 15'000.–
    - de plus de Fr. 15'000.–
    - pas de réponse
50. Quel est le numéro postal d'acheminement de votre domicile ?

Nous sommes ainsi parvenus à la conclusion de notre interview. Merci bien. Il peut arriver que nous appelions une fois brièvement pour une question complémentaire ou une demande de précisions. Mais ce n'est que rarement nécessaire. Nous vous remercions vivement pour vos renseignements précieux et vous souhaitons une excellente soirée.

D Fehlende Werte im Referenzmodell

Tabelle D.1: Fehlende Werte im Referenzmodell

Variable	Anz. fehlende Werte
f85300	682
f03301	498
f03302	499
f03303	496
f03304	494
f03305	493
f03306	493
f03307	499
f03308	494
f03309	502
f03310	493
f03311	498
f03312	493
f03313	492
f03314	495
f03315	493
f03401	498
f03402	494
f03403	493
f03404	495
f03405	493
f03406	519
f03407	494
f03408	492

Fortsetzung nächste Seite . . .

Tabelle D.1: Fehlende Werte im Referenzmodell

Variable	Anz. fehlende Werte
f03409	499
f03410	493
f03501	488
f03502	488
f03503	488
f03504	488
f03505	488
f85100	492
N85200	645

E Verwendete Software

Statistik Alle Berechnungen, Modelle und Simulationen wurden mit der Open Source Programmiersprache R entwickelt und gerechnet (R Development Core Team, 2010). Es kamen die Versionen 2.9.3 bis 2.12.0 zum Einsatz Innerhalb von R wurden verschiedene *packages* verwendet.

- Für die Berechnung der Bäume wurde das Packet `rpart` (Therneau et al., 2009) und zur Kontrolle `tree` (Ripley, 2009) eingesetzt.
- Grafiken wurden vorwiegend mit den Packeten `ggplot2` (Wickham, 2009) und `Lattice` (Sarkar, 2008) erstellt.
- Neben den Funktionen aus den Basispacketen von R, waren für die Berechnung der Regressionen folgende Packete hilfreich: `car` (Fox und Weisberg, 2010), `lmtest` (Zeileis und Hothorn, 2002) und `arm` (Gelman et al., 2010).

Textsatz Die Erstellung des Textdokuments erfolgte in LaTeX (Lamport, 1994). Dabei war insbesondere das KOMA-Script hilfreich (Kohm und Morawski, 2008). Als Schriftarten wurden Linux Libertine (Textschrift) und Linux Biolinum (serifenlos) verwendet.

Editor Als Editor für LaTeX diente Vim (Robbins et al., 2008). Der Editor für R war Eclipse, Version 3.6.1 Helios (`www.eclipse.org/documentation/`) mit dem StatET-Package (`http://www.walware.de/goto/statet`). Auch Vim wurde als Editor für R verwendet (mit Vim-R-Plugin, `http://www.vim.org/scripts/script.php?script_id=2628`).

Betriebssystem Der Haupteil der Arbeit wurde auf einem Rechner mit Mac OS X v.10.6 «Snow Leopard» geleistet. Für einige Berechnungen und Kontrollen wurde auch ein Windows-System verwendet (Windows XP

SP3). Die Simulationen wurden auf einem Server mit dem Betriebssystem Linux Red Hat durchgeführt.

F R-Code

Um den Verlauf der Simulationen besser verstehen zu können, wird der R-Code für die Simulation mit GREG-Schätzer (Variante Referenz) aufgeführt. Siehe dazu auch Abschnitt 10.3, ab S. 196. (Für die Schätzung des GREG wurden zwei verschiedene Packages ausprobiert: `sampling` (Tillé und Matei, 2009) und `survey` (Lumley, 2010). Im untenstehenden Code wird `sampling` verwendet.

```
1  ##########################
2  #Hilfreiche Funktionen definieren
3
4  #quintile berechnen
5  quin <- function(vec,sep=5) {
6          neu <- sort(vec)
7          out <- numeric(sep-1)
8          for (i in 1:(sep-1)) {
9                  out[i]<-neu[(i*(round((length(neu)/sep),0)))]
10         }
11         out
12 }
13 #Gewichtete Anteile im LaTex Tabellen Format
14 wtab <- function(var,wgt) {
15         var <- as.factor(var)
16
17         levels(var)<- c(levels(var),"99")
18         var[is.na(var)]<- "99"
19         ltx <- character()
20         out <- numeric(length(levels(var)))
21         for (i in 1: length(levels(var))){
22                 out[i]<-round(sum(wgt[var==levels(var)[i]])/sum(wgt)*
                       100,2)
23         }
24         ltx <- as.character(out[1])
25         for (i in 2:length(levels(var))){
26                 ltx <- paste(ltx," & ", out[i],sep="")
27         }
28         ltx <- paste(ltx,"\\",sep="")
29         ltx
30 }
```

```
#Gruppen bilden ps: Variable zum Gruppieren, gru: Referenzgruppe gru
    ==1, l=Anz Quantile
grup <- function(ps,gru,l=5){
        qui <- quin(ps[gru==1],l)
        gr <- rep(l,length(ps))
        for (i in l:2) gr[ps<qui[i-1]]<-(i-1)
        gr
}
grup.w <- function(gr,gru){
        all <- table(gr)
        pan <- table(gr[gru==1])
        w <- numeric(length(gr))
        for (i in l:length(gr)){
                wi <- (all[i]/sum(all))/(pan[i]/sum(pan))
                w[gr==i]<-wi
        }
        w
}
#Informationskriterien Regressionen
ic <- function(link.model){

#Informationskriterien Probit Modell
        probit.np<-length(link.model$coefficients)
        probit.ns<-length(link.model$residuals)

#probit.McFR2
        probit.McFR2<-1-link.model$deviance/link.model$null.deviance

#probit.CSR2
        probit.CSR2<-1-exp(-link.model$null.deviance/2+link.model$
            deviance/2)^(2/probit.ns)

#probit.AIC und BIC
        probit.npar<-length(link.model$coef)
        probit.AIC<-link.model$deviance+2*probit.npar
        probit.BIC<-link.model$deviance+log(probit.np)*probit.npar

#probit.NR2
        probit.R2max<-1-exp(-link.model$null.deviance/2)^(2/probit.ns)
        probit.NR2<-probit.CSR2/probit.R2max

#probit.vif funktioniert nicht
        probit.vif <- vif(link.model)

#probit.kappa
        probit.kappa <- kappa(link.model)


        probit.ic <- numeric()
```

```
78          probit.ic[1]<-round(probit.McFR2,2)
79          probit.ic[2]<-round(probit.CSR2,2)
80          probit.ic[4]<-round(probit.AIC,1)
81          probit.ic[3]<-round(probit.NR2,2)
82          probit.ic[5]<-round(probit.BIC,2)
83          probit.ic[6]<-round(probit.kappa,1)
84          names(probit.ic)<-c("McFadden R2", "Cox-Snell-R2", "Nagelkerke
                R2", "AIC", "BIC/SBC", "Kappa")
85          probit.ic
86  }
87
88  ##############################################################
89  #Daten lesen und Variablen definieren
90  link <- read.csv("/Workspace/Ablage/linkR.csv")
91  link <- link[(link$f85100!=2 & link$f85100!=99),] # Non - Web User
        raus
92  link <- link[link$f85401!=1,] #MArktforscher raus 17
93  link <- link[!is.na(link$f03001),] #14 Beobachtungen mit vielen
        fehlenden Werten raus
94
95  #Boost raus
96  link <- link[link$sample==1,]
97  link$N85200[is.na(link$N85200)]<-0  #186 fehlende Werte Anz
        OnlinekÃ¤ufe
98  #Anzahl fehlender Werte zÃ¤hlen
99  link$noNA <- 0
100 for (i in 1: ncol(link)){
101         link$noNA <- link$noNA + as.numeric(is.na(link[,i]))}
102 #Missings sind entweder 98,99 oder 999
103 missing.list <- names(link)
104 for (i in 1:length(missing.list)) {
105         eval(parse(text = paste("link$",missing.list[1],"[link$",
                missing.list[1],"==98] <- NA", sep="")))
106         eval(parse(text = paste("link$",missing.list[1],"[link$",
                missing.list[1],"==99] <- NA", sep="")))
107         eval(parse(text = paste("link$",missing.list[1],"[link$",
                missing.list[1],"==999] <- NA", sep="")))
108 }
109 #kategorische Variablen
110 als.kategorisch <- c("status", "f00110", "f00140", "HHgroesse", "
        f85300","f03305","f03306", "f90300", "f91100", "f91600", "f03301"
        , "f03302", "f03307", "f03308",  "f03311", "f03312", "f03401", "
        f03402", "f03403", "f03404", "f03405", "f03407", "f03408", "
        f03409", "f03410", "f03501", "f03502", "f03503", "f03504", "
        f03505", "f85100", "f86110", "f03303", "f03304", "f03309","f03310
        ","f03311","f03312","f03313","f03314","f03315", "f03406" ,"f03405
        ", "f03405", "f03405", "f03200", "N90401", "N90402","N90403", "
        N90404","N90405","N90406", "f91300", "anzErw")
111
```

```
112 for (i in 1: length(als.kategorisch)) {
113         eval(parse(text = paste("link$",als.kategorisch[i],"  <- as.
                factor(link$",als.kategorisch[i],")", sep=""")))
114 }
115 #Variablen konstruieren
116 link$x1 <- factor(nrow(link),levels=c(1,2,3))
117 link$f03501<-as.numeric(link$f03501)
118 link$x1[link$f03501<4]<- 1
119 link$x1[link$f03501==4]<- 2
120 link$x1[link$f03501>4]<- 3
121 link$x2 <- factor(nrow(link),levels=c(1,2,3))
122 link$f03502<-as.numeric(link$f03502)
123 link$x2[link$f03502<3]<- 1
124 link$x2[link$f03502==3]<- 2
125 link$x2[link$f03502>3]<- 3
126 link$x3 <- factor(nrow(link),levels=c(1,2,3))
127 link$f03503<-as.numeric(link$f03503)
128 link$x3[link$f03503<3]<- 1
129 link$x3[link$f03503==3]<- 2
130 link$x3[link$f03503>3]<- 3
131 link$x4r <- as.numeric(link$f03505)+rnorm(nrow(link))+2*as.numeric(
      link$x2)
132 link$x5r <- as.numeric(link$f03306)+rnorm(nrow(link))+2*as.numeric(
      link$x2)
133 link$x6r <- as.numeric(link$f03305)+rnorm(nrow(link))+2*as.numeric(
      link$x2)
134 #abhÃŒngige Variable y
135 link$y <- numeric(nrow(link))
136 link$y <- 5*as.numeric(link$x1) + (as.numeric(link$x2)+3)^2 +
137                 5*as.numeric(link$x3) + link$u +
138                 (link$x4r + link$x5r + link$x6r)/2
139 ##########################################
140 #Packages lesen
141 library(sampling)
142 #Anzahl Simulationsrunden
143 runden<-10000
144 #Output vorbereiten
145 out<-matrix(nrow=runden,ncol=8)
146 #Seed= 0 wegen Reproduzierbarkeit
147 set.seed(0)
148 link$u <- rnorm(nrow(link))
149 #totals and design.matrix for greg
150 tdat <- model.matrix(~link$x2)
151 tdat<-as.data.frame(tdat)
152 colnames(tdat)<-c("a","b","c")
153 tdat$a[(tdat$b+tdat$c)==1]<-0
154 tdat <- cbind(tdat,link$x4r,link$x5r,link$x6r)
155 iw <- rep(1,nrow(tdat))
156 ss <- c(sum(as.numeric(tdat$a)), sum(as.numeric(tdat$b)),
```

```
157                 sum(as.numeric(tdat$c)),sum(link$x4r),sum(link$x5r),
                        sum(link$x6r))
158 ############################################################
159 #Simulationsschlaufe
160 for (j in 1:runden){
161 link$u <- rnorm(nrow(link))
162 #responsewahrscheinlichkeit
163 link$R0 <- numeric(nrow(link))
164 link$R0 <- as.numeric(link$x2)^2 + 2*as.numeric(link$x3) +
165                 link$u - (link$x4r + link$x5r + link$x6r)/2.6
166 link$r <- exp(link$R0)/(1+exp(link$R0))
167 #Stichprobe ziehen, n=200 Einladungen
168 link$R <- logical(nrow(link))
169 link$R[sample(1:nrow(link),200,prob=link$r)]<-T
170 #Modell
171 link.model<-glm(R~x2+x4r+x5r+x6r, family=binomial(link="logit"),data=
        link)
172 #Informationskriterien Probit Modell
173 probit.np<-length(link.model$coefficients)
174 probit.ns<-length(link.model$residuals)
175 #probit.CSR2
176 probit.CSR2<-1-exp((-link.model$null.deviance+link.model$deviance)/
        probit.ns)
177 #probit.NR2
178 probit.R2max<-1-exp(-link.model$null.deviance/probit.ns)
179 probit.NR2<-probit.CSR2/probit.R2max
180 #predict PS
181 link$ps <- predict(link.model,type="response")
182 link$wps <- 1/link$ps
183 #Klassifikation
184 q.ps <- quantile(link$ps[link$R],seq(0.2,1,0.2))
185         #Einteilung in Quintile der PS im Sample
186         link$ps.k<-rep(1,nrow(link))
187         for (ii in 1:length(q.ps)-1){
188                 link$ps.k[link$ps>q.ps[ii]]<-ii+1
189         }
190         #Berechnung Gewichte
191         link$w.k <- rep(NA,nrow(link))
192         for (ii in 1:length(q.ps)){
193                 link$w.k[link$ps.k==ii] <- table(link$ps.k)[ii]/table(
                        link$ps.k[link$R])[ii]
194         }
195 #GREG
196 link$wg <- rep(0,nrow(link))
197 link$wg[link$R] <- calib(Xs=tdat[link$R,],d=iw[link$R],total=ss,method
        ="linear")
198 #Outputs
199 m.ps <- weighted.mean(link$y[link$R],link$wps[link$R])
200 v.ps <- max(link$wps)/min(link$wps)
```

```
m.k <- weighted.mean(link$y[link$R],link$w.k[link$R])
v.k <- max(link$w.k)/min(link$w.k)
m.g <- weighted.mean(link$y[link$R],link$wg[link$R])
ms      <- mean(link$y[link$R])
mgg <- mean(link$y)
nr2 <- probit.NR2
out[j,] <- c(mgg,ms,nr2,m.ps,v.ps,m.k,v.k,m.g)
}
out <- as.data.frame(out)
colnames(out) <- c("gg","s","nr2","ps","vps","k","vk","g")
```

G Lebenslauf

Ich wurde am 7. August 1975 in Eisenach, damals DDR, geboren. Meine Eltern sind Carola und Werner Wiegand. Umzugsbedingt habe ich viele verschiedene Schulen besucht: eingeschult wurde ich als Siebenjähriger in Treffurt, habe dann an die Botschaftsschule in Bogotá (Kolumbien) gewechselt und nach erneuten Zwischenstationen in Suhl und Eisenach am «Albert-Schweizer-Gymnasium» in Ruhla 1994 mein Abitur mit den Leistungsfächern Mathematik und Geschichte abgelegt.

Nach meinem einjährigen Zivildienst bei den Behindertenwerkstätten des Diakonieverbunds Eisenach habe ich 1995 begonnen, Soziologie mit den Nebenfächern Kunstgeschichte und Logik zu studieren. Nach einem Jahr Studium habe ich allerdings vom Magister- zum Diplomstudium gewechselt und meine ursprünglichen Nebenfächer zugunsten der Volkswirtschaftslehre aufgegeben. Im Studienjahr 1997/98 habe ich das Studium in Leipzig unterbrochen, um ein Jahr an der University of Teesside in Middlesbrough Economics zu studieren. 2001/02 habe ich das Studium in Leipzig mit der Erlangung des Titels Diplom-Soziologe abgeschlossen. Das Thema der Diplomarbeit war «Die Entstehung vertrauensbildender Normen», betreut von Prof. Dr. Thomas Voss.

Während und nach dem Studium habe ich bei der GIABmbH in Erfurt im Bereich empirischer Sozialforschung gearbeitet. Bei der GIAB handelte es sich um einen Spin-Off der Universität Erfurt.

2004 habe ich Doris Wiegand, geb. Blaser, geheiratet und bin zu ihr in die Schweiz gezogen. Am 2. November 2007 wurde unsere Tochter Linda geboren, am 15. Juni 2011 unsere Tochter Paula. Seit dem Frühjahr 2011 besitze ich neben der deutschen auch die Schweizer StaatsbÃ¼rgerschaft.

Seit 2005 arbeite ich bei der Fachhochschule Nordwestschweiz an der Hochschule für Wirtschaft im Schwerpunkt von Prof. Dr. Beat Hulliger. Zusammen mit Beat Hulliger habe ich das KTI-geförderte Projekt «Online Panel Qualität» bearbeitet, welches die dieser Arbeit zugrunde liegende Befragung

ermöglicht hat. Für die vielen wichtigen Anregungen bedanke ich mich bei Beat Hulliger herzlich.

Die mündliche Prüfung zu dieser Dissertation erfolgt im Fach Soziologie am 17. August 2011 in Basel.